ELSEVIER
JAI

ADVANCES IN LEARNING AND BEHAVIORAL DISABILITIES

VOLUME 19

APPLICATIONS OF RESEARCH METHODOLOGY

THOMAS E. SCRUGGS
MARGO A. MASTROPIERI
Editors

APPLICATIONS OF RESEARCH METHODOLOGY

ADVANCES IN LEARNING AND BEHAVIORAL DISABILITIES

Series Editors: Thomas E. Scruggs and
Margo A. Mastropieri

Recent Volumes:

Volume 14: Edited by Thomas E. Scruggs and Margo A. Mastropieri

Volume 15: Edited by Thomas E. Scruggs and Margo A. Mastropieri

Volume 16: Edited by Thomas E. Scruggs and Margo A. Mastropieri

Volume 17: Edited by Thomas E. Scruggs and Margo A. Mastropieri

Volume 18: Edited by Thomas E. Scruggs and Margo A. Mastropieri

ADVANCES IN LEARNING AND BEHAVIORAL DISABILITIES VOLUME 19

APPLICATIONS OF RESEARCH METHODOLOGY

EDITED BY

THOMAS E. SCRUGGS
George Mason University, Fairfax, USA

MARGO A. MASTROPIERI
George Mason University, Fairfax, USA

ELSEVIER
JAI

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo
JAI Press is an imprint of Elsevier

JAI Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

First edition 2006

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN-13: 978-0-7623-1295-5
ISBN-10: 0-7623-1295-5
ISSN: 0735-004X (Series)

For information on all JAI Press publications
visit our website at books.elsevier.com

Printed and bound in The Netherlands

06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

CONTENTS

LIST OF CONTRIBUTORS

Stephanie Al Otaiba	College of Education and Florida Center for Reading Research, Florida State University, USA
Giulia Balboni	University of Valle d'Aosta, Italy
Jeanine Clancy-Menchetti	Florida Center for Reading Research, USA
Maureen Conroy	Department of Special Education, University of Florida, USA
Samantha Dietz	College of Education, University of Miami, USA
Dimiter Dimitrov	College of Education and Human Development, George Mason University, USA
Jennifer Dixon	College of Education, University of Miami, USA
Elizabeth A. Edgemon	Curry School of Education, University of Virginia, USA
Craig Enders	Department of Psychology, Arizona State University, USA
Rosa García	Department of Developmental, Educational and Social Psychology and Methodology, University Jaume I of Castellón, Spain
Marona Amandla Leaura Graham-Bailey	Peabody College, Vanderbilt University, USA
Brian R. Jablonski	Curry School of Education, University of Virginia, USA
Olga Jerman	Graduate School of Education, University of California, Riverside, USA

Kathleen Lane	Peabody College, Vanderbilt University, USA
John W. Lloyd	Curry School of Education, University of Virginia, USA
Margo A. Mastropieri	College of Education and Human Development, George Mason University, USA
Kimberly A. McDuffie	Clemson University, USA
Marjorie Montague	College of Education, University of Miami, USA
Kelley S. Regan	Department of Special Education, George Washington University, Washington, DC, USA
E. Jemma Robertson	Peabody College, Vanderbilt University, USA
Edward J. Sabornie	College of Education, North Carolina State University, USA
Christopher Schatschneider	Department of Psychology and Florida Center for Reading Research, Florida State University, USA
Thomas E. Scruggs	College of Education and Human Development, George Mason University, USA
Manuel Soriano	Department of Developmental and Educational Psychology, University of Valencia, Spain
Janine P. Stichter	Department of Special Education, University of Missouri, Columbia, USA
Lee Swanson	Graduate School of Education, University of California, Riverside, USA
Andrew L. Wiley	Center for Social Development and Education, University of Massachusetts, Boston, USA

QUALITATIVE RESEARCH APPLICATIONS WITH YOUTH WITH HIGH-INCIDENCE DISABILITIES

Edward J. Sabornie

ABSTRACT

The contributions of qualitative research to the study of behavioral–emotional disabilities, mild intellectual disabilities, and learning disabilities (the three types of high-incidence disabilities) are relatively recent and far from abundant. This chapter discusses qualitative, or "naturalistic" research by briefly examining the methodology used in such inquiry, reviewing many of the available studies concerning those with high-incidence disabilities, and providing implications from the existing empirical literature. It is not recommended that qualitative research takes the place of quantitative research in special education, but well-designed and executed naturalistic studies can contribute additional knowledge that is worthwhile to the field.

The impact of qualitative research in special education, in comparison with group or single-subject design studies, is a relatively recent phenomenon. While qualitative inquiry has established a slight foothold in special education, and it appears to be growing in popularity among doctoral students

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 1–32

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19001-1

involved with disability-oriented dissertations, it is still far from the major means of empirical production in the field. Perhaps it is because the majority of special education researchers are not captivated by its contributions, or that most investigators are simply uninformed in using such methodology. Qualitative research has also been somewhat shunned in special education because it has been closely associated with postmodern philosophy, and some leaders in "naturalistic," or "interpretive" research (e.g., Lincoln & Guba, 1985) believed that it should not be forced to co-exist with quantitative research. Perhaps this latter mentality has trickled down to the masses.

Skrtic (1986) and Stainback and Stainback (1984) served as catalysts for increased interest in "multiparadigmatic" research in special education, although their conclusions regarding the beauty and power of qualitative inquiry were not universally accepted at the time (see Kauffman, 1987; Simpson & Eaves, 1985; Ulman & Rosenberg, 1986). According to Stainback and Stainback, the potential of qualitative research lies in its ability to expand our viewpoint and awareness of students who are identified as exceptional. Without the knowledge that naturalistic research can add to our understanding – at least in the eyes of the Stainbacks and Skrtic – the field of special education is left with an incomplete picture of those with disabling conditions, how they are treated in school and served in the community, what challenges they face in life, and especially how they think and feel.

Many qualitative studies related to those with high-incidence disabilities (i.e., behavioral–emotional disabilities (BED), mild intellectual disabilities (MID), and learning disabilities (LD)) typically include vibrant descriptions of participants and their dialog, and carefully drawn images of the contexts in which phenomena exist that interact with persons experiencing disabilities. The qualitative researcher interviews participants to gain additional personal perspective from an "expert" on the matter under study. The weight assigned to qualitative interviews is usually judged by the trustworthiness of the data related to confirmability, credibility, dependability, and transferability (Guba, 1981). Naturalistic inquiry seeks to "ground" its research focus on a variable examined in an investigation and, in so doing, completely highlight the issue, with imperfections included, through interpretation provided by the researcher. According to some (see Kavale & Forness, 1998; Simpson & Eaves, 1985), it is in the interpretative nature of knowledge production where qualitative research is weakest regarding objectivity and science.

Some qualitative research designs (see below) depend on consistent and multiple direct observations of behavior and its context. In some types of naturalistic studies the researcher or "inquirer" must completely immerse himself or herself into an environment as a "participant observer" so that all aspects of an environment or phenomenon are felt first hand. For that reason the social context is very important in many qualitative studies. Participant observation is also aimed at providing greater potency to the interpretation provided by the researcher. Some interpretive research uses *triangulation*, which provides for different interpretations of a phenomenon, using many sources of information, and various qualitative data collectors. Analysis of interpretive data can be made easier through computer programs such as *NUD*IST* and *The Ethnograph*, among others. "Member checks," in which a data collector asks participants to double-check field notes and a researcher's interpretation of a respondent's statement, are also found in many qualitative studies (cf., reliability checks in single-subject research). In other words, while Lincoln and Guba (1985) believed that qualitative research need not compete with or co-exist with quantitative research, the two types of inquiry share more characteristics than some would care to admit.

The purpose of this chapter is to examine published qualitative research as it applies to students and people with high-incidence disabilities. A précis of interpretive research methodologies will be provided as well as a review of selected qualitative research studies concerning persons and issues related to those with BED, MID, and LD. Implications from the reviewed naturalistic research will also be offered.

QUALITATIVE RESEARCH METHODOLOGIES

The five most frequently used designs or "traditions" of qualitative research include: (a) ethnography, (b) grounded theory, (c) biographical, (d) case study, and (e) phenomenological (Creswell, 1998). Following is a brief description of the methods used with each type of naturalistic inquiry.

Ethnographic Research

Because of its extensive use in anthropology by Mead (1963), ethnographic research is probably the best known of all the types of naturalistic research. The origin of the participant observer used in many types of interpretive analysis can be traced to ethnographic research, and there are many types of qualitative scholarship that include variations of the theme and methods found in ethnographic inquiry. Those who espouse philosophies such as postmodernism and Marxism, among other schools of thought, have also adapted and used ethnography to highlight their beliefs (Tedlock, 2003).

The *sine qua non* of traditional ethnographic inquiry is the immersion of the researcher in coterie. What ethnographers attempt to provide is a comprehensive interpretation of all the nuances, ethnicity, and life cycles of a *culture* so that the reader understands the group ''from the inside.'' The data collector uses field notes and observes people going about their daily tasks and rituals, interviews as many participants as possible from all walks of life to gain an extensive array of perspectives, and analyzes such data to expose the culture or group with all its complexity. This type of research is far from easy or ephemeral for in order to uncover all that is endemic to a culture or group requires a considerable time commitment on the part of the ethnographic investigator. Todis, Bullis, Waintrup, Schultz, and D'Ambrosio (2001), for example, studied the life histories of 15 adolescents with behavior and emotional problems over the course of five years (see below). What emerges from an ethnographer's published field notes is a holistic representation of a group with all its successes, problems, and peculiarities.

Grounded Theory Research

Qualitative research based on grounded theory was originally developed in the 1960s by sociologists Glaser and Strauss (1967). In grounded theory inquiry, a researcher formulates questions about a phenomenon, collects data on the item of interest, analyzes the preliminary data and reconstructs the phenomenon, then collects additional data and reconstructs until satisfied and a tentative theory emerges. This iterative process is used to ensure that any theory constructed from the available data is an accurate description of the phenomenon under study. Logistically, it is perhaps one of the more difficult types of qualitative research because of the back-and-forth construction and reconstruction process involved in this type of inquiry.

Grounded theory inquiry requires the researcher to interview participants that are closely involved with a phenomenon (e.g., social skills of adolescents with high-incidence disabilities; see Kolb & Hanley-Maxwell, 2003, below). Such interview data form *categories* that are descriptive of the phenomenon under consideration. Along with the interviews, the researcher may also choose to observe participants interacting with the phenomenon, if possible. At some point in time after interviews have been conducted the researcher begins data analysis that leads to the formulation of preliminary theoretical constructs; subsequent data collection (e.g., interviewing) continues until the research categories are *saturated* and the phenomenon is completely uncovered in the mind of the researcher. The formation of categories that may change over time with additional data collection and saturation is referred to as "constant comparative" data analysis (Creswell, 1998).

A standard sequence of research steps is followed in grounded theory data analysis. *Open coding* is the first stage whereby data categories and subcategories are formed with the initial data. The next phase of analysis involves *axial coding* in which new categories are formed in addition to what was established in the open coding phase, and the specific context and consequences of the phenomenon are exposed. *Selective coding* follows next whereby categories and subcategories are commingled and the investigator forms a tentative hypothesis concerning the phenomenon. A research report traditionally follows selective coding which attempts to explain the phenomenon in narrative form with numerous excerpts from interviews supporting any hypotheses generated. Students learn in introductory qualitative inquiry courses that any grounded theory researcher is duty-bound to curb subjective views of the construct under examination so that an untainted perspective emerges from a systematic process. Suppression of subjectivity on the part of the researcher, however, is difficult at best because of the personal (i.e., non-software generated) manner in which categories and subcategories may emerge from the multi-step coding process.

Biographical Research

Biographical qualitative research has existed in disability-related interest areas at least since the days of the noted physician, Itard, and the classic examination of his patient, Victor, in the *Wild Boy of Aveyron* (Itard, 1962). Traditional naturalistic inquiry of a biographical nature consists of one person telling a researcher about his or her labors and life experiences, and the investigator then brings such stories to text. Typical of biographical research is a seminal event in a person's life that serves as an axis for personal change, or change of perspective, and this incident is interpreted by the qualitative investigator. Different types of biographical inquiry include oral histories, autobiographies, life histories, and individual biographies (Creswell, 1998).

To begin this type of research, the biographical inquirer explores the available written documents and records of the participant of interest and describes the person in terms of the various stages of life experienced. Written records of another are interpreted and brought to life in the narrative, and the investigator also describes the relationship that he or she has formed with the main character of the biographical study. The majority of the qualitative data presented in a biographic study, however, are the conversations and interviews between the participant and biographer. When the participant relates that a noteworthy event led to a change of direction or perspective, the researcher typically visits the important place and attempts to comprehend the context of the environment, and describes it in the narrative so that any consumer of the research understands its significance. In other words, the researcher attempts to interpret the critical event for the reader and includes the investigator's own impressions of the incident and its context. The investigator also connects the participant's turning point to the larger world at the time of the incident as well as present circumstances, and generally completes the biography with "lessons learned" (Lincoln & Guba, 1985).

Case Study Research

Case study naturalistic research is often associated with investigations of individuals, but a "case" need not be concerned only with people, but an activity, program, or event that is bounded by place and time (Creswell, 1998; Stake, 1994, 1995). Case studies can examine (a) different types of the same phenomenon (e.g., a few community colleges serving students with LD), (b) one very unique case or intervention (e.g., a special, 4-year college serving only students with LD and those with attention deficit hyperactivity disorder such as Landmark College in Vermont), or (c) one example from a set of similar cases (e.g., the services a "Students with Disabilities Program" office offers at a large, state-supported university).

Researcher and participant observations, interviews of individuals involved with a phenomenon, and archival record examinations are all characteristics of the methodology of case study qualitative research. Also found in this type of research is *purposeful sampling*, the converse of random sampling. The purposeful sampling procedure requires choosing a variety of cases (or informants) that present different views of the same construct in an open-ended data collection time frame. The tactics used in case study research can vary from when a researcher attempts to expose an entire phenomenon (e.g., the efforts of all the general educators serving students with BED in an elementary school), to *embedded analysis*, where a very specific trait of a case is uncovered (e.g., teaching reading to elementary level students with MID in general education classes). Similar to other types of qualitative inquiry, in case study research the inquirer uses extensive field notes and anecdotal records which comprehensively describe the phenomenon and its context and document the chronology of events that touch the case. The investigator should also include his or her relationship and past history with the case in the narrative. Data analysis involves a broad search for themes and subthemes and couching the results in the social context where the case resides.

Phenomenological Research

Perhaps the most controversial of all the forms of qualitative inquiry in the eyes of the positivistic researcher, phenomenological methodology relies on heavy interpretive techniques and postmodern philosophies (Sabornie, 2004). The following captures the philosophy that undergirds this type of inquiry: "Phenomenologists reject the scientific *realism* (emphasis in original) and the accompanying view that the empirical sciences have a privileged position in identifying and explaining features of a mind independent world" (Schwandt, 2001, p. 191). In this type of research, the investigator describes the subjective statements of participants and structures and derives meaning from such information. Contrary to a quantitative researcher involved in analyzing graphs, behavioral trends, and statistical data, a phenomenologist instead describes subjective feelings and emotions of participants who have interacted in some manner with the same phenomenon. The actual personal reactions expressed by participants are the data source, and the goal of the researcher is to uncover the underlying structure to all the sentiments.

A typical phenomenological study opens with the researcher stating his or hers views toward the construct under study, and how experience has colored the researcher's views toward it. The narrative of this type of inquiry also includes an attempt by the researcher to see the phenomenon through the eyes of the participants, and to derive new or different meaning from the view of the contributors. The inquirer attempts to experience the construct in the same way as the participants, and conducts interviews with participants before, during, and after contact with the phenomenon. Important participant statements (in the eyes of the researcher) describing the phenomenon are chosen and analyzed to determine if themes emerge to indicate some new or uniform structure to the comments. In a phenomenological narrative the researcher organizes the discourse assuming that participants' subjective impressions of a construct, event, or incident have a specific structure, and the configuration of the contributors' emotions is revealed.

Similar to the variety and combinations of research designs shown in quantitative and behavioral inquiry, the five types of qualitative research designs discussed above have many variations (e.g., action research, critical theory research, discourse analysis, focus group research, narrative research, quasi-life-history research; see Brantlinger, Klinger, & Richardson, 2005). One should also keep in mind that the origin of the five qualitative traditions was not in education but in other social sciences such as sociology, psychology, anthropology, and even political science (Creswell, 1998).

Except in studies involving issues pertinent to LD, there is very little subject matter pattern to the extant interpretive research involving youth with high-incidence disabilities. In light of this shortcoming, the studies concerning issues in BED and MID that are reviewed below are presented without regard to a specific subject matter (i.e., dependent variable); discourse related to excellence of design and execution, and magnitude of contribution of the studies, is found below in the conclusions section. Last, many studies included in the review did not specifically examine actual students or persons with high-incidence disabilities but rather chose to examine issues and phenomena related to the participants with BED, MID, and LD.

RESEARCH CONCERNING BEHAVIORAL–EMOTIONAL DISABILITIES

Because the source of most special education inquiry is in applied behavior analysis and quantitative, group design statistics, the amount of qualitative research involving any one group of students with disabilities is not voluminous. This is especially true of research involving students with BED, and most of the available naturalistic studies originate somewhat recently in the 1990s and early 2000s.

Crowley (1993) used an ethnographic design to examine the perceptions of six adolescents with BED toward helpful and harmful teacher behaviors in general education classes. The participants had a history of aggression in school, and they spent at least one period per day in a general education classroom. The author conducted between four and eight 50-min interviews of each participant over a period of six months. Crowley asked the adolescents specific questions during the audiotaped interviews (e.g., "What kinds of things do teachers do that you find helpful in your general education classes?"), and also used direct observation of the participants (i.e., in 36, 50-min sessions) in the general education settings. Borrowing from applied behavior analysis, Crowley also performed interrater reliability checks of the observational data collected.

Six themes emerged from the extensive data collection: Three (i.e., teacher–student communication, flexible academic programming, and flexible behavioral programming) were related to accommodating teacher behavior, two (i.e., punitive disciple practices and teacher rigidity) were aligned with non-supportive teacher interactions with the students, and the last theme was associated with students' anger. The anger theme was subdivided into four subthemes: (a) nondescript anger, (b) anger toward classmates, (c), teacher-directed anger, and (d) anger toward the school administration. Crowley (1993) concluded that naturalistic research (in combination with some applied behavior analysis research methods) can add much to our understanding of interventions meant to assist students with BED.

Todis et al. (2001) used ethnographic research methods over five years to highlight resilience among 15 formerly incarcerated adolescents. The researchers used audiotaped interviews of the 15 actual participants, and also 44 family members, teachers, and friends of the youth to gain a complete picture of the adolescents of interest in the study. Participant observation was also used in homes, places of work, and other settings that the participants visited during the first year of the study. The interviews and observation field notes were reduced to form categories and subcategories involving relationships among the respondents and other tentative theories. Many life histories were presented through vivid narratives of the participants' experiences from the predelinquent years, the incarcerated period, and up to current status. Todis et al. were able to categorize the life histories of the participants in three groups: "succeeders" ($n = 6$), "drifters" ($n = 7$), and "strugglers" ($n = 2$). A statement from a Gary, a succeeder, included: "I just want to see how far I can get in the world" (p. 135), and Sally, a drifter stated, "I haven't been doing too good as a teenager" (p. 135). The researchers concluded that parents, schools, and other adult support services need to improve if we expect formerly incarcerated adolescents to become succeeders. One of the hallmarks of qualitative research is its almost proud disregard for the notion of generalizability (and psychometric reliability and validity, too; see Janesick, 1994) of findings to a larger population, yet Todis et al. used their provincial results to address the needs of a larger group not specifically examined in their study.

Kolb and Hanley-Maxwell (2003) used grounded theory research to examine parents' perspectives concerning the social skills of their middle school offspring with the three types of high-incidence disabilities. The purpose of the study was to determine what parents of adolescents with high-incidence disabilities thought about social skills, and what the parents wanted their offspring to learn in school in the social domain. Data and parents' responses were not separated by category of exceptionality, and the 11 youth whose parents' responses were examined ranged in age from 12 to

15. Interviews, phone conversations, informal chats, and extensive field notes written subsequent to interactions with the parents were used to gather the perspectives of the participants, and open, axial, and selective coding along with triangulation were also part of the methodology. What emerged from the data was that parents want their adolescent youth to learn about social skills and be successful in the social domain. Parents expect the teaching of social skills to be immersed into academic instruction so that their offspring can obtain social intuition and empathy. The authors also discussed the long-standing problem concerning the time involved to teach social skills, versus the time required to teach academic content in the present day of high stakes testing for adolescents with disabilities. It appears that the time necessary to teach social skills to students with disabilities is a problem no matter what research methodology one uses to examine it.

Epstein and Quinn (1996) used case study methods to uncover the experiences of a 15-year-old boy with behavior disorders within a system of care. The purpose of the study was to shed light on the relationship between the child, the family, and the system of care for children and youth with BED. The participant had a history of difficulties in school (e.g., aggressive behavior, not completing assignments, theft, arson), and he had been identified as BED since the elementary school level. Data collected included information from archival records, and participant and parent interviews. Timelines were constructed which included the chronology of the participant's life events and involvement with the local and state care agencies, as well as the cost involved in serving the adolescent. Over time, the participant had received care in a state-run mental hospital, spent two years at an out-of-state residential facility, was placed in an alternative school while living at home, and also resided in a local residential care facility. Results showed that the system of care was not very successful in improving the participant's problem behaviors, and that the youth did not demonstrate much interest regarding planning for his own treatment. The total cost of care for the participant exceeded an astounding $290,000 with little, if any, improvement shown in his conduct disorders. Although endemic to only one participant, the Epstein and Quinn case study is a good example of how naturalistic inquiry can add to our understanding of students with disabilities in ways not typically found in quantitative and applied behavior analysis research.

Morningstar, Turnbull, and Turnbull (1995) used purposeful sampling along with focus group methodologies to examine adolescents' with high-incidence disabilities ideas concerning the importance of family involvement in the transition process from school to adult life. The participants included in the four focus groups were 13 adolescents (ages 13–19) with BED, 9 with

MID, and 18 with LD; responses were not separated by category of disability. Focus group members were interviewed and a moderator guided the discussion following open-ended questions. Discussion categories were formed from within and across focus group interviews. Morningstar et al. were able to collapse the focus group responses into three domains: vision of the future, stakeholder involvement in the transition process, and facilitation of self-determination. Family guidance was evident for the youth in their movement from school to independent adult life, and some careers chosen by the focus group members followed other family members' wishes. Most participants were not actively involved in planning for the future and did not attend their own Individual Education Plan (IEP) meetings because they found such gatherings meaningless. Some focus group members were allowed autonomy in personal decisions while others looked to family members for assistance. The Morningstar et al. study was able to shed light on the personal perspectives of some adolescents with high-incidence disabilities, which would not ordinarily be found in other types of research.

In another study using focus groups over a one-year period, Lehman and Fredericks (1997) investigated the characteristics of service coordination for families of children with BED. Eight parents of children with BED and six professionals involved in service coordination were involved as focus group participants. Data collection consisted of a 2-h focus group meeting and in-depth interviews of the participants. The focus group meeting data were analyzed before the interviews so that the entire scope of service coordination would be uncovered by at least one type of data collected. Member checks were used for scrutinizing response authenticity along with two-source (i.e., parents, service coordinators) triangulation to gain perspective on the phenomenon under study. Results showed that three themes materialized concerning effective service coordination for students with BED: (a) characteristics of the service provider organization, (b) personal characteristics of the actual service provider interacting with the family, and (c) global characteristics of the community and service subsystems. Genuine concern for the child and family, and flexible availability time were found to be helpful to parents involved with seeking and receiving aid from service coordinators. Some of the barriers to effective service provision included blaming and judging the parents, and unfamiliarity with current best practices.

Summary. Given the very few qualitative studies found in peer-review journals to this date, it is safe to say that this type of research has not found its place yet in the study of BED (Sabornie, 2004). Even though naturalistic research in special education has been defended more than once since the mid-1980s (see Crowley, 1994–1995; MacArthur, 2003; Peck & Furman, 1992; Scruggs & Mastropieri, 1995; Stainback & Stainback, 1984), the justifications for such inquiry have barely touched researchers and most of the important educational issues surrounding BED. The available published studies have also not focused on successful school- or community-based interventions. It is hoped this will change over time for seminal contributions similar to that of Crowley (1993) and Epstein and Quinn (1996) can add greatly to the extant knowledge concerning BED.

RESEARCH ON ISSUES RELATED TO MILD INTELLECTUAL DISABILITIES

Page and Chadsey-Rusch (1995) used a variation of case study methodology to expose the experiences of four youth attending community colleges, two of whom had intellectual disability. Data originated from six full-day interviews with the participants, observations of the four youth while moving about campus, and other documents concerning the students from the community college and their former public schools. Extensive field notes, audiotaped interviews, data triangulation, member checks, additional auditing of the data by another researcher, and unspecified "attempts to obtain evidence contrary to the emerging findings" (p. 88) were used as methods to ensure accuracy of interpretation and findings. Two large themes emerged from the data: (a) the decision to attend community college (subthemes: expectations of others and built-in supports) and (b) impact of attending community college (subthemes: future career plans, and interpersonal experiences). Enrolling in postsecondary education was an expectation expressed by the two students who were nondisabled, but was not true of the participants with intellectual disability. The support system found at the community college allowed the students with intellectual disability to participate and find varying levels of success in the postsecondary educational environment. Contrary to the courses the nondisabled participants chose at the community college, the two with intellectual disabilities enrolled in courses without reference to a future job. Last, the participants with intellectual disabilities mentioned meeting new people and feeling "grown-up" because of the community college experience. The authors concluded that more than anything else, the two participants with intellectual disability achieved interpersonal benefits from attending community college. While postsecondary level education is often recommended in transition-related objectives on IEPs for adolescents with LD, it appears that students with intellectual disability can also find it useful and fulfilling.

Mactavish, Mahon, and Lutfiyya (2000) used individual interviews, focus groups, and additional verification meetings to examine social integration from the perspective of 32 persons (aged 17–82-years old) with varying levels of severity of intellectual disability. Fourteen of the participants had MID; participant responses were not grouped by level of severity of disability. A secondary goal of the study was to illuminate the research strategies that enabled the participants with intellectual disabilities to actively share their views with the researchers concerning social integration. Individual interviews of the participants were conducted to build rapport, collect background information, and explain the goals of the research to respondents. Eight focus groups were then formed to explore the participants' many perspectives concerning social integration. After the focus group meetings the researchers and participants attended verification conferences so that interpretations of the respondents' dialog could be checked and refined. The results showed that the 32 participants viewed social integration as "the sense of belonging that emanated from the sharing of time, activities, and experiences with families and friends – independent of whether these individuals had a disability" (p. 224). An interesting caveat was that the research team erroneously assumed that social integration would involve the interactions between people with and without disabilities. Contexts that served as mechanisms for social integration included school, work, family, day activity programs, and living arrangements. These same catalysts created both positive *and negative* perceptions of social integration among the participants. The researchers concluded that if social integration is to be fully understood the subjective views of participants with disabilities must be represented.

Devlieger and Trach (1999) used ethnographic research with extensive life history interviews to evaluate the effect of mediation on employment and transition outcomes of six adolescents with MID. In an effort to expose the research context of each participant, the researchers constructed extensive social networks of the six youths, and comprehensive descriptions of their personal characteristics are found in the narrative of the study. The researchers were particularly interested in how each participant secured a specific employment or living situation, and whether there was involvement on the part of adult agencies or "self-family-friends" that led directly to various work or living arrangements. Data collection procedures included (a) interview sessions with the participants and family members, (b) review of agency documents related to each adolescent, and (c) field notes of informal interactions such as visiting the person at home or having dinner with them at a fast-food restaurant. "Scripts," or a "composite picture that describes the process of transition from school to adult life" (p. 516) were developed for each person based on all the qualitative information uncovered in the data collection. Results showed that each participant had a very unique individual script with specific mediators that contributed to his or her transition success. One participant without an extensive social network or mediators experienced little success in post-school employment. The authors concluded that students with MID need someone or an agency to serve as a mediator and problem solver to help guide adolescents through independent adult life.

Zetlin (1986) used participant observation over 18 months to study the relationships of 35 adults (18 males and 17 females) with MID and their siblings. The mean IQ level for the group with MID was 67 (range 60–69), with a mean age of 34 (range 23–60). The main methodologies used for data collection included monthly meetings that typically lasted several hours and regular phone calls with the participants. Structured life histories of the participants were also constructed via interviews with close family members. Field notes were examined to determine references by the participants to siblings, and data were coded into the categories of warmth, frequency of contact, and degree of involvement (i.e., dependency and reciprocity). Spanning a continuum from very close to uninvolved, five levels of connection emerged from the data that characterized the relationships among adults with MID and their siblings: (a) very warm feelings with frequent contact and involvement; (b) warm feelings with regular contact and moderate involvement; (c) warm feelings with minimal contact and minimal involvement; (d) resentful feelings with minimal contact and involvement; and (e) hostile feelings with rare or no contact or involvement. The largest single type of relationship mentioned by 20 participants and 27 sibling pairs was one with warm feelings and minimal contact and involvement. Another finding was that existing sibling relationships tended to be hierarchical with nondisabled relatives frequently providing assistance and support for their brothers and sisters with MID. Sisters provided greater support and care for their siblings with MID than did brothers. Zetlin found it somewhat surprising to uncover such great variation in sibling relationships.

Brantlinger (1988) used an "emergent naturalistic research design" to examine teachers' perceptions of the sexuality of their adolescent students with MID. The researcher purposefully chose 22 teachers of secondary level students with MID to form the research sample, and a flexible, open-ended set of questions related to sexual knowledge and behaviors were posed to the respondents. Field notes were written while interviewing the secondary level special education teachers along with audiotaping the sessions; interviews lasted up to 2 h. Member checks were also performed with four teachers to assess for authenticity and plausibility of data and interpretation. Results showed that, in general, teachers did not feel that their students with MID acquired accurate information concerning sexual topics. Two subgroups of students emerged from the interviews – "streetwise" and "naïve" – based largely on socioeconomic status (SES). Adolescents with MID from lower SES were found most often in the streetwise subgroup and thought to be more active at an earlier age and to a greater extent than were the naïve. Nearly all teacher-participants commented that their students thought sex was dirty and associated it with shame. Teachers also felt that students were opposed to abortion and surrendering babies for adoption. Many teachers also served students with odd sexually oriented behaviors and had experience with pregnant students in their classes. Two-thirds of the teachers felt that pregnancy was somewhat desired by their students. Brantlinger concluded that teachers need a comprehensive sexuality education curriculum to better educate students with MID in this important domain.

Boyce, Marshall, and Peters (1999) used qualitative diary analyses to document the stressors, coping responses, and "uplifts" of six adolescents with disabilities. Two participants were diagnosed with intellectual disability, two with Down syndrome, one had cerebral palsy, and one had a learning disability, but findings and conclusions were not separated by category of disability. The authors created a semistructured daily journal for use by the participants which consisted of open ended queries that required respondents to answer "This is what stressed me today … ," "This is what I did about it … ," and "The best part of today was … " The daily diary entries were subjected to content analysis to identify categories and themes that were shown in the qualitative data. Two researchers reviewed all data categorizations to ensure similarity of opinion across data coders. The researchers gathered a total of 97 diary entries (range = 7–32, mean = 16), 30 from girls and 67 from boys. Results showed that of the 97 journal notes, about 20% did not include a stressor. Three types of stressors were shown by the participants: (a) environmental situations; (b) cognitive, emotional, and physical aspects of self; and (c) personal relationships. Coping responses were grouped into the categories of cognitive-behavioral (e.g., "I cried"), cognitive-intrapsychic (e.g., "I remember that other people do the same thing."), and interpersonal responses ("I talked to my teacher about it."). The uplifts noted by the participants included four categories: social-physical activities, personal satisfaction, comforting situations, and personal relationships. The Boyce et al. study is yet another example of how qualitative research can illuminate the perspectives of persons with disabilities and, in so doing, increases the level of understanding we have of another's life experiences.

Hagner and Davies (2002) conducted an interesting case study of self-employed adults with intellectual disability in New Hampshire, Vermont, and Massachusetts. The purpose of the study was to broadly describe the types of employment owned and operated my adults with cognitive disabilities, and to uncover levels of support and quality of work life for the eight business owners. Seven participants were described as having cognitive disabilities with an additional business owner having traumatic brain injury. Four in-depth interviews (between 30 and 90 min) with the participants, and observations of the business were used for data collection. An individual closely associated with providing employment support for the business owner was also interviewed. Some of the businesses included the production and sale of women's jewelry, decorative gift baskets, painted wooden figures and letters, and in-home childcare.

Results showed that most business owners with a cognitive disability operated their enterprise only on a part-time basis. A lack of suitable other employment opportunities was an important factor in owning a private business for five of the eight participants. The actual businesses developed around the participants' own values and interests, and start-up funds usually originated with the person's family or support network. Four of the persons with cognitive disabilities received formal business training from staff members of a developmental services agency, and one woman entrepreneur (the childcare business owner) received financial assistance from a local businesswomen's organization. Unfortunately, most of the businesses provided very little income for the owners (e.g., $30 to $50 gross sales for the gift basket business), and several businesses required subsidies to "break even." Moreover, many support personnel stated that time involvement and level of assistance necessary to help the business operate was substantial and beyond what was typically expected. While one must give credit to the adults for starting and operating their own company, it appears that self-employment of these specific participants with cognitive disabilities is difficult at best, and not very lucrative.

Stainton and Besser (1998) used grounded theory narrative data, collected using a quasi-focus group methodology, to examine the impact that children with an intellectual disability have on families. Nine family units were observed in 2-h semistructured group interviews, and two additional interviews were conducted with two separate families who had a child with intellectual disability. The intellectual disability of the children ranged from mild to severe, but responses were not assembled by level of severity of disability.

Using a constant comparative method the researchers identified general themes related to quality of impact on the family of the child with intellectual disability. Reliability checks on the narrative data were also performed by an independent researcher so that inconsistencies and omissions could be identified. Results showed that nine themes emerged from the interviews. Seven themes were related to the positive impact on the family of having a child with intellectual disability (e.g., source of joy and happiness, increased sense of purpose and priorities, source of family unity), and two external to family impact were also uncovered (e.g., interaction with professionals and services, and having a positive impact on others outside the family and in the community). The findings confirm those of other researchers (e.g., Turnbull, Guess, & Turnbull, 1988) showing that having a child with intellectual disability is not filled with only pain and sorrow.

Bigby (1998) examined how 62 adults with intellectual disability, aged 55 and older, used community services in a grounded theory qualitative investigation. The sample consisted of adults with intellectual disability at the mild or moderate level of severity, and the participants' responses were not categorized by level of disability. Data were collected via semistructured interviews with the adults with intellectual disability, service providers, and an informant (i.e., siblings, friends, cousins) who had a long-term relationship with the participant with intellectual disability. The purpose of the interview was to gain perspective concerning the participants' lives and the informal and formal service support available and used in the community. Constant comparative methods were used to generate the themes found in the interview data. Results showed, surprisingly, that the large majority (85%) of participants used age-related services more so than disability-oriented support. Twenty-one percent of the participants used both age- and disability-care services. The interview data revealed that the types of services most commonly used were residential primary care from a group home, supported employment, leisure activities, case management, and intermittent household management supervision. The Bigby study also shows that older persons with intellectual disability need generic services not specifically related to any disability.

Summary. The quantity of studies related to issues and persons with MID reviewed above ($n = 9$) again shows that qualitative inquiry is far from a dominant source of empirical knowledge for this particular group. In fact, with the exception of the Brantlinger (1988) and Page and Chadsey-Rusch (1995), the remaining seven studies reviewed do not focus on actual elementary or secondary level public school *students* with intellectual disability. Even Brantlinger did not specifically examine students with MID, but instead polled their teachers, and Page and Chadsey-Rusch examined community college students.

Perhaps this lack of qualitative research quantity is yet another indicator of the waning interest associated with MID. "The field of mild mental retardation is becoming endangered, with little current research and questions from the field if it is time for the population's eulogy" (Bouck, 2005, p. 309). Many other published qualitative studies exist (e.g., Cambridge & Forrester-Jones, 2003; Medved & Brockmeier, 2004; Patching & Watson, 1993; Richardson, Kline, & Huber, 1996, to name just a few) that examined issues pertaining to those with more moderate to severe levels of intellectual disability. Those with intellectual disabilities comprise the third largest group of students with disabilities in the United States – 10.6% of all students aged 6 through 21 receiving special education services – and roughly 85% of such individuals fall into the mild range of severity (Beirne-Smith, Patton, & Kim, 2006). That equals over 500,000 students. It appears that without additional and continued research of any type the field is eschewing an important and sizeable population, and part of the long history of special education in the U.S.

RESEARCH ON MATTERS RELATED TO LEARNING DISABILITIES

Of the three groups with high-incidence disabilities, those with LD are the largest in terms of number of students and with regard to the quantity of qualitative research involving them as participants. MacArthur (2003) reported that between the years 1991 and 2001, 82 qualitative studies dealing with issues concerning LD were published in four of the leading journals in special education and LD. Space restrictions do not allow for the review of all the studies published concerning every issue related to LD, but those included for review below represent a cross-section of the available qualitative inquiry in the field.

Adults with Learning Disabilities

Reiff and colleagues (i.e., Reiff, Gerber, & Ginsberg, 1993, 1994; Reiff, Ginsberg, & Gerber, 1995; Shessel & Reiff, 1999) have provided four studies in which the perspectives and experiences of adults with LD were examined. In Reiff et al. (1995), the researchers used a grounded theory approach to describe the paths that 71 persons with LD traveled to become successful adults. The sample of adults included 41 in the "highly successful" subgroup, and 25 in the "moderately successful" cluster; the mean age of the participants was in the mid-40s, and income levels ranged from less than $10,000 to over $1,000,000 per year. Quantitative data were also collected on the participants which included scores on a self-esteem measure, a vocational achievement motivation questionnaire, and a workplace relationships scale. All participants were interviewed using a protocol of 130 open-ended questions across nine categories such as success, family, social issues, and daily living. The interviews with each participant lasted from 3 to 8 h with a mean length of 4.5 h. Results were divided into internal and external issues that led to the success of the adults with LD. Internal traits included having control, a desire to succeed, goal orientation, and awareness of having a learning disability. External manifestations of participants' success were shown in persistence, being able to maximize strengths and minimize weaknesses, and having creativity in approaching tasks and problems. Most of the successful adults with LD also mentioned that support from families and seeking help for specific situations were important contributors to their accomplishments. The authors concluded that teachers of students with LD should avoid focusing on the failures and attempt to discover the possibilities of success for their pupils.

Shessel and Reiff (1999) used ethnographic interviewing procedures to examine the positive and negative impacts and outcomes in the lives of 14 adults with LD. Participants ranged in age from 26 to 60 with educational levels from 11th grade to a master's degree. The participants were interviewed twice with each interview lasting about 60–90 min in length. The audiotaped interviews consisted of participants answering open-ended questions concerning family and educational history, social-emotional functioning, vocational experiences, and daily living issues. Member checks were conducted during the second interview. Results were reported with reference to the positive and negative effects of living with a LD. Four themes were associated with the negative impacts and outcomes: daily living issues (e.g., sometimes the disability caused interference and frustration), the "imposter phenomenon" (e.g., "pretending to be something that you're not"), social isolation and perception (e.g., feeling socially isolated and "being different," and that their LD interfered with their social lives), and emotional health (e.g., stress, anxiety, negative self-concept, fear of "looking stupid"). The positive aspects of having an LD included helping the participants to be a better person, allowing them to think creatively, developing a desire to help others, making them better professionals, and increasing their sensitivity to others. The participants in this study wished to make their feelings known so that others with LD would have a reasonable quality of life and personal fulfillment.

The findings concerning adult participants in Shessel and Reiff (1999), especially those of a negative nature in the social domain, are very similar to those expressed by middle-school adolescents with LD in another qualitative study that examined the outcomes of being labeled LD (see Reid & Button, 1995). Additional personal opinions of what it is like to have a LD, drawn from naturalistic research and from students of various ages, can be found in Albinger (1995), Barga (1996), and Guterman (1995).

Bilingual Instructional Issues in Learning Disabilities

Ruiz (1995a) used ethnographic techniques over 20 months to examine bilingual Latino students, aged 6–11, identified as "language learning disabled" (LLD). The intent of the study was to reveal the impact of context on the students' language and academic functioning and to create performance profiles of groups of students that indicated actual functioning. The study took place in a self-contained, elementary level, bilingual special education classroom. Ruiz acted as a participant observer in the classroom for 16 months where she observed (using field notes) and audiotaped the teacher, paraprofessional, and students engaged in their normal classroom routines. During the last four months of data collection the researcher reviewed the students' IEPs, cumulative school files, and interviewed the parents of the participants. Three different student performance profiles, ranging from severe language disability to normal functioning, emerged among the student participants. Four children were found in the moderate to severe language disability group, three children were found in the mildly language disabled to normal functioning group, and two students were categorized as having normal ability. Ruiz questioned why the two students who were perceived to be normal were actually educated in a self-contained classroom, and she used the results to condemn the medical, or deficit model in special education.

Ruiz (1995b), using the same participants, teachers, and classroom as in Ruiz (1995a), again used ethnographic procedures to discover the types of "classroom events" that characterize the language usage among bilingual students with language LD. The researcher used 28 entire school day observations of the self-contained classroom environment to collect 32 h of audiotaped classroom activities along with comprehensive field notes. Ruiz found that three types of classroom events were largely responsible for language activities, and that these events could be classified along a continuum from most formal (class opening), to moderate formality (academic lessons), to least formal in scope (sociodramatic play). She also found that the classroom event that allowed for the exhibition of the upper range of students' language ability was the least formal, or sociodramatic play. In essence, the more formal the classroom event, the more the students displayed language difficulties. Even in the sociodramatic play activity, however, students with the most severe LD displayed less frequent language participation and more errors in their verbal message. Among other conclusions related to language acquisition, Ruiz used the results to disparage the behaviorist or "reductionist" model of verbal communication instruction used extensively in special education. Additional qualitative research related to bilingual instructional issues in special education can be found in Ruiz, Rueda, Figueroa, and Boothroyd (1995), and Echevarria and McDonough (1995).

Science Instructional Issues in Learning Disabilities

Scruggs and Mastropieri (1994a) used qualitative methods to study instruction and inclusion of students with disabilities. The purpose of the investigation was to expose factors associated with successful inclusion of students with disabilities in science class across grade levels and across four different types of disabilities. Nine students with LD participated, along with students with hearing, visual, and physical disabilities; qualitative data were not divided into areas related to each disability category. The researchers observed and videotaped three science classrooms in grades three through five during two consecutive school terms. A total of 50 class meetings were observed. Students, teachers, and administrators were also interviewed and students' and teachers' work products were collected. Triangulation of the data was used in that multiple sources of evidence were found to support the conclusions. An additional, and very robust data treatment exercise was also used by the researchers in that all final conclusions were examined to ensure that each was supported by qualitative evidence collected in the observations, interviews, and work products. Results showed that the three classrooms were successful in including the students with four different types of disabilities into science class, and that seven variables were linked with the success of the inclusion program. These variables were: (a) administrative support; (b) support from special education personnel; (c) an accepting, positive classroom climate; (d) appropriate curriculum; (e) effective general teaching skills; (f) peer assistance; and (g) disability-specific teaching skills. Students with LD were able to use grade-level science materials and curricula when the emphasis on reading and writing was slight. This study provides confirmation that certain elementary level students with LD and other types of disabilities can succeed in grade-level content courses if the right instructional variables are in place.

In an "analytic induction" investigation involving academic instruction, Scruggs and Mastropieri (1994b) examined the scientific reasoning of students with LD and MID in elementary level classes over a two-year period. Eight students with LD and six with MID, across grades one through five, served as participants. The students met together two days per week with special education teachers and paraprofessionals to receive science instruction. Classroom observations along with interviews of the principal, teachers, and students served as the source of the naturalistic data collected. Six major themes emerged concerning how students with high-incidence disabilities created scientific knowledge, which included:

- students were able to construct scientific knowledge with scientific methods;
- effective teaching behaviors were related to student construction of scientific knowledge;
- adaptations of the science curricula were made to meet the special learning needs of the students;
- teacher-implemented behavioral techniques maintained appropriate behavior, attention, and consistent effort;
- successful knowledge construction by the student required highly structured coaching by the teachers; and
- peers assisted with social encouragement and skill applications, but were not as skilled in encouraging learning outcomes.

The authors concluded that when students with high-incidence disabilities use reason to understand scientific content they are likely to remember and use such information. An additional qualitative study involving students with LD in inclusive science classes is found in Palincsar, Magnusson, Collins, and Cutter (2001).

Literacy Instructional Issues in Learning Disabilities

Vaughn, Moody, and Schumm (1998) used both qualitative and quantitative measures to examine the reading instructional and grouping practices of 14 special education resource room teachers. The elementary level teachers served 77 students with LD, two with MID, and three with orthopedic disabilities. Vaughn et al. interviewed the teachers at the beginning and end of the school year and used open-ended queries about the teachers' backgrounds, grouping and reading instructional practices, and their views concerning effective reading curricula. Each interview lasted between 30 and 60 min. The 14 resource room teachers were observed three times during the school year during reading and language arts instructional periods. In addition, on the three days when classroom observations occurred the teachers were asked to complete a checklist that pertained to the actual instruction observed (e.g., "What was the composition of your groups?" "Who selected the materials that were used in the different groups?"). The data presented by the researchers were organized into seven themes: grouping practices, individualized instruction, overall approach to reading instruction, teaching word recognition and decoding, teaching comprehension, monitoring student progress in reading, and teachers' views of special education. Among the numerous findings, Vaughn et al. found that reading instructional grouping practices usually involved three or four different grade levels, only a few of the teachers provided differential work for the students, 10 of the 14 teachers stated that whole language was the primary approach used to teach reading, and only three of the teachers provided ongoing word recognition and decoding instruction. The students of the 14 teachers made little or no progress during the one-year length of the study. Vaughn et al. concluded that not enough students in resource rooms received the proper and most effective methods of reading instruction.

Moody, Vaughn, Hughes, and Fisher (2000) conducted a two-year follow-up investigation of some of the same resource room teachers in the original Vaughn et al. (1998) study discussed above. Similar interview and observation ($n = 4$) techniques were used in the follow-up research to examine six of the 14 teachers in the original study. The student population in the 6 resource rooms consisted of 59 students with LD, and four others identified as having either MID, BED, autism, or orthopedic disability. The foci of the Moody et al. follow-up was to determine if any instructional changes had occurred in some of the resource rooms in light of reading reform initiatives concerning the teaching of explicit skills and phonics. Results showed that the six teachers were much more concerned with the teaching of phonics in comparison to the prior two years. The increased concern about the teaching of phonics, however, did not translate into an increased level of the actual teaching of phonics in the resource rooms. Grouping practices in the follow-up study changed in that the six teacher-participants used (a) less whole group instruction, (b) more individualized instruction, and (c) instructional groups in which students were at the same reading level. One-half of the teachers in the follow-up study used differentiated materials and instruction to match the students' reading levels. Many of the follow-up teachers stated that their view of special education was to address the social, emotional, and self-confidence needs of students in the resource room. Similar to the results of the original study, students in the Moody et al. follow-up did not gain significantly in reading comprehension over the course of one school year. The researchers concluded that the resource room teachers were not solely to blame for the lack of reading progress made by the students. Teachers' caseloads were so large that effective instruction and individualized reading attention were nearly impossible to deliver.

Additional qualitative research concerned with literacy among students with LD is found in Englert, Berry, and Dunsmore (2001), Englert, Rozendal, and Mariage (1994), MacArthur, Schwartz, Graham, Molloy, and Harris (1996), Mariage (1995, 2000, 2001), and Palincsar, Parecki, and McPhail (1995).

Full Inclusion Issues in Learning Disabilities

In one of the most far-reaching qualitative investigations of students with high-incidence disabilities in special education to date, Baker and Zigmond (1995; also see entire issue of *The Journal of Special Education*, *29*(2), 1995) used five different case studies – in five states – to document full inclusion school services delivered to students with LD. The larger purpose of the collection of case studies in Kansas, Minnesota, Pennsylvania, Virginia, and Washington was to determine whether students with LD in inclusive classrooms were receiving a *special* education. Ten students with LD in six school buildings in elementary level and intermediate grades served as participants. Each pair of participants in each state was observed in school over a two-day period during academic instruction in general education settings. Data collectors used field notes to document the behavior of teachers, class members, and target students with LD. Semistructured interviews were conducted with the participants, parents of the child, general and special education teachers, school principal, and building special education supervisors. Students' records were also examined to gather data on achievement levels and IEPs of the participants. Member checks were performed before case summaries were compiled.

Themes were constructed around features found across the five cases and included model of inclusion, role of the special education teacher, and educational events of the participants with LD. Results showed that the cases differed widely in terms of who led the inclusion program delivered to the students. Some were developed and directed by university leaders, and two were locally driven. Students with LD were included in the classrooms of only those teachers who volunteered to be part of the inclusion team. In some cases six to eight students with LD were found in one classroom, while in others a disproportionate number of pupils with LD were not found in any one classroom. Various models of instructional delivery were found in the general education classrooms including consultative, co-teaching, peer and paraprofessional assistance, and services provided beyond the classroom and traditional school day (e.g., extended school day programs). All the special education teachers spent time in the inclusion classrooms, and their activities ranged from teaching only students with IEPs to delivering instruction to a whole class. The typical arrangement was when the students were divided into two instructional groups, and the special and general education teachers each taught a separate group in the same classroom. Students with LD experienced modified assignments and materials, and the most common accommodation was when instruction was changed for the entire group to meet the needs of the students with LD. At the same time, however, some general education teachers were opposed to making any accommodations for the students with LD. Remedial instruction in reading and math was available to some participants in extended school day activities. Teachers attempted to address the special needs of students with LD by providing peer and paraprofessional assisted instruction. The researchers concluded that the participants with LD received a very good *general* education, students experienced insufficient "specially designed instruction," and that the inclusion models required more resources than did traditional pull-out special education.

Using qualitative research methods, Pugach and Wesson (1995) examined a full inclusion educational program from the perspective of nine students with LD and three teachers (two general, one special education) in two fifth-grade classrooms. The participants with LD were fully integrated into the two classrooms for an entire school year, and the total number of students in both classrooms equaled 55. Prior to the study, the students with LD were educated in a pull-out resource room program for part of the school day. Data sources included the transcripts of interviews with nine nondisabled students and the nine participants with LD, and open-ended interviews with the three teachers. The student interviews were 20–45 min in length, and the teacher interview was an open-ended meeting of 2 h. Classroom observations of the participants and teachers were not used. Results presented three themes (i.e., classroom social climate, instructional effects, teacher roles and tasks) and 10 subthemes from the interviews. On the basis of only interviews, the researchers found a positive social climate in the classroom where the both the students with LD and nondisabled peers felt good about their teachers and themselves. Regarding the instruction in the classrooms, the majority of the participants with LD and all nine of the nondisabled pupils enjoyed the variety of instructional activities. Grouping practices in the fifth-grade classrooms were flexible, and special assistance was delivered individually or in small groups. The students did not view the special education teacher as attending to students with learning difficulties exclusively, and they felt that their social and academic needs were being met better than in a single classroom with only one teacher. Pugach and Wesson concluded that it is possible for full inclusion classrooms to promote feelings of success for both students with LD and the nondisabled.

Additional qualitative studies involving students with LD and full inclusion school environments include Cutter, Palincsar, and Magnusson (2002) and Rice and Zigmond (2000). Related naturalistic research concerning students with LD and the "coaching" and consultation process that occurs among special and regular education teachers involved with full inclusion include Gersten, Morvant, and Brengelman (1995) and Marks and Gersten (1998).

Summary. The quantity of naturalistic research involving issues of importance in LD is likely due to the size of the population – those with LD comprise the single largest group of students aged 6–21 in the public schools of the U.S. Even with the lack of concern for generalization found in qualitative inquiry, the available studies with issues pertaining to LD, unlike similar research involving matters and students with BED and MID, also allows for the examination of important issues in the field across several topical studies. Moreover, one study related to LD reviewed herein (e.g., Scruggs & Mastropieri, 1994a) should serve as the model for subsequent interpretive inquiry in special education because of its rigorous data treatment methodology. If additional studies employed the same data handling as that found in Scruggs and Mastropieri perhaps qualitative research would not be frowned upon by so many positivist paradigm investigators.

CONCLUSIONS

Qualitative inquiry exists as a nontraditional approach to research that adds subjective understanding to contexts, constructs, and populations. While scattered in foci except for studies in LD, qualitative research concerning those with high-incidence disabilities is now viewed as acceptable inquiry. More studies need to be conducted, however, that would shed additional light on what it is like to have MID or BED (cf. Albinger, 1995; Barga, 1996). Cross-categorical qualitative studies that hunt for similitude and differences across students with BED, MID, and LD are also needed. A recent meta-analysis and a descriptive review of studies found students in the three categories of high-incidence disabilities are more dissimilar than alike (see Sabornie, Evans, & Cullinan, in press; Sabornie, Cullinan, Osborne, & Brock, 2005). It would be interesting to determine if qualitative research examining cross-categorical issues mirrored the same findings. What would also be helpful in qualitative research is a method to aggregate findings across studies similar to what is done with meta-analysis in group designs (see Scruggs, Mastropieri, & McDuffie, this volume). Meta-analysis partials-out subjectivity in interpreting a great number of studies on the same dependent variable, but interpreter bias can still enter into any descriptive review of a body of qualitative research.

There is a need for additional research related to successful academic interventions in special education (Trout, Nordness, Pierce, & Epstein, 2003). Qualitative research can show what is academically effective at the local level with provincial context as the background, and large-scale quantitative research can address the scalability issue by showing that educational interventions can be generalized to a larger population. In this way both types of research can co-exist and contribute meaningfully even though it may oppose the opinions of some experts in qualitative research (see Lincoln & Guba, 1985).

Qualitative research should never be considered as a substitute for positivist inquiry for the old adage “there is no such thing as a perfect study” still holds no matter what design methodology is used. As Kauffman (1987) stated regarding naturalistic research, “none of these methods and no combination of them is a sufficient replacement for quantitative analyses” (p. 61). Because of the dependence on human interpretation in much of qualitative research, one still needs to view naturalistic inquiry through the lens of its shortcomings. Humans make mistakes, and misinterpretations can happen without member checks, triangulations of data, and when participant interviews are not paired with observations.

A continuing threat to the existence of qualitative research in special education concerns the importance of results. Because statistics are not an issue in naturalistic inquiry one still needs to judge the findings of a study based on *practical* significance. Beyond the reader’s impressions, judging a qualitative study with regard to practical significance is difficult at best. Local knowledge of a phenomenon gleaned from qualitative research has some merit, but the question still remains in terms of *how much* value. Some (e.g., Lincoln & Guba, 1985) have said that you cannot judge a qualitative study using the same metrics as those used to evaluate a group design investigation, but to ignore the practical significance of a study is analogous to increasing ambiguity. Adding uncertainty to a subject matter or variable is the opposite of what research on any theme, using any type of methodology, should accomplish.

While qualitative and quantitative research have specific strengths, it may come as a surprise to some that both approaches share some design and execution weaknesses. Participant sampling procedures plague both types of research, and duration of data collection, choice of instrumentation, categorizing of data, and choosing the best design based on type of data collected confound both paradigms. In addition, it is just as easy to interpret beyond the findings and data in quantitative research as it is to do so in interpretive inquiry.

If for no other reason, qualitative research in special education should be congratulated for dispelling some myths associated with "the Other." The concept of "Otherness" is related to viewing persons with disabilities (and other people who do not make up the majority of a race, culture, or other population) as different based solely on the presence of a disability (Bogdan & Knoll, 1995; Murdick, Shore, Chittooran, & Gartin, 2004). Qualitative inquiry has opened the doors and given the narratives of persons with disabilities a place in the extant literature that was previously denied to them. Naturalistic research has also shown that persons with disabilities have some of the same wants, needs, and desires that are typical among the nondisabled. Putting a more human face on persons with disabilities is a proper outcome for any type of research to pursue in special education.

REFERENCES

Albinger, P. (1995). Stories from the resource room: Piano lessons, imaginary illness, and broken-down cars. *Journal of Learning Disabilities*, *28*, 615–621.

Baker, J. M., & Zigmond, N. (1995). The meaning and practice of inclusion for students with learning disabilities: Themes and implications from the five cases. *The Journal of Special Education*, *29*, 163–180.

Barga, N. K. (1996). Students with learning disabilities in education: Managing a disability. *Journal of Learning Disabilities*, *29*, 413–421.

Beirne-Smith, M., Patton, J. R., & Kim, S. H. (2006). *Mental retardation* (7th ed.). Upper Saddle River, NJ: Pearson Merrill.

Bigby, C. (1998). Shifting responsibilities: The patterns of formal service use by older people with intellectual disability in Victoria. *Journal of Intellectual and Developmental Disability*, *23*(3), 229–243.

Bogdan, R., & Knoll, J. (1995). The sociology of disability. In: E. A. Meyen & T. M. Skrtic (Eds), *Special education and student disability* (pp. 667–711). Denver, CO: Love.

Bouck, E. C. (2005). Impact of factors on curriculum and instructional environments for secondary students with mild mental retardation. *Education and Training in Developmental Disabilities*, *40*, 309–319.

Boyce, G. C., Marshall, E. S., & Peters, M. (1999). Daily stressors, coping responses, and uplifts of adolescents with disabilities. *Education and Training in Mental Retardation and Developmental Disabilities*, *34*, 406–417.

Brantlinger, E. (1988). Teachers' perceptions of the sexuality of their secondary students with mild mental retardation. *Education and Training in Mental Retardation*, *23*, 24–37.

Brantlinger, E., Klinger, J., & Richardson, V. (2005). Importance of experimental as well as empirical qualitative studies in special education. *Mental Retardation*, *43*, 92–119.

Cambridge, P., & Forrester-Jones, R. (2003). Using individualized communication for interviewing people with intellectual disability: A case study of user-friendly research. *Journal of Intellectual & Developmental Disability*, *28*(1), 5–23.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Crowley, E. P. (1993). A qualitative analysis of mainstreamed behaviorally disordered aggressive adolescents' perceptions of helpful and unhelpful teacher attitudes and behaviors. *Exceptionality*, *4*, 131–151.

Crowley, E. P. (1994–1995). Using qualitative methods in special education research. *Exceptionality*, *5*, 55–69.

Cutter, J., Palincsar, A. S., & Magnusson, S. J. (2002). Supporting inclusion through case-based vignette conversations. *Learning Disabilities Research & Practice*, *17*, 186–200.

Devlieger, P. J., & Trach, J. S. (1999). Mediation as a transition process: The impact on postschool employment outcomes. *Exceptional Children*, *65*, 507–523.

Echevarria, J., & McDonough, R. (1995). An alternative reading approach: Instructional conversations in a bilingual special education setting. *Learning Disabilities Research & Practice*, *10*, 108–119.

Englert, C. S., Berry, R., & Dunsmore, K. (2001). A case study of the apprenticeship process: Another perspective on the apprentice and scaffolding metaphor. *Journal of Learning Disabilities*, *34*, 152–171.

Englert, C. S., Rozendal, M. S., & Mariage, M. (1994). Fostering the search for understanding: A teacher's strategies for leading cognitive development in "zones of proximal development." *Learning Disability Quarterly*, *17*, 187–204.

Epstein, M. H., & Quinn, K. P. (1996). A case study approach to analyzing the relationship between children and services in a system of care. *Journal of Emotional and Behavioral Disorders*, *4*, 21–29.

Gersten, R., Morvant, M., & Brengelman, S. (1995). Close to the classroom is close to the bone: Coaching as a means to translate research into classroom practice. *Exceptional Children*, *62*, 52–66.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago: Aldine.

Guba, E. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal*, *29*, 75–92.

Guterman, B. R. (1995). The validity of categorical learning disabilities services: The consumer's view. *Exceptional Children*, *62*, 111–124.

Hagner, D., & Davies, T. (2002). "Doing my own thing": Supported self-employment for individuals with cognitive disabilities. *Journal of Vocational Rehabilitation*, *17*, 65–74.

Itard, J. M. G. (1962). *The wild boy of Aveyron* (G. Humphrey, & M. Humphrey, Trans.). New York: Prentice-Hall.

Janesick, V. J. (1994). The dance of qualitative research design. In: N. K. Denzin & Y. S. Lincoln (Eds), *Handbook of qualitative research* (pp. 209–219). Thousand Oaks, CA: Sage.

Kauffman, J. M. (1987). Research in special education: A commentary. *Remedial and Special Education*, *8*(6), 57–62.

Kavale, K. A., & Forness, S. R. (1998). The politics of learning disabilities. *Learning Disability Quarterly*, *21*, 245–273.

Kolb, S. M., & Hanley-Maxwell, C. (2003). Critical social skills for adolescents with high incidence disabilities. *Exceptional Children*, *69*, 163–179.

Lehman, C. M., & Fredericks, H. D. (1997). *Qualitative investigation of effective service co-ordination for children and youth with emotional and behavioral disorders (Report No. EC 305831)*. Monmouth, OR: Western Oregon State College (ERIC Document Reproduction Service No. ED411640).

Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

MacArthur, C. (2003). What have we learned about learning disabilities from qualitative research? A review of studies. In: H. L. Swanson, K. R. Harris & S. Graham (Eds), *Handbook of learning disabilities* (pp. 532–549). New York: Guilford.

MacArthur, C. A., Schwartz, S. S., Graham, S., Molloy, D., & Harris, K. (1996). Integration of strategy instruction into a whole language classroom: A case study. *Learning Disabilities Research & Practice*, *11*, 168–176.

Mactavish, J. B., Mahon, M. J., & Lutfiyya, Z. M. (2000). "I can speak for myself": Involving individuals with intellectual disabilities as research participants. *Mental Retardation*, *38*, 216–217.

Mariage, T. V. (1995). Why students learn: The nature of teacher talk during reading. *Learning Disability Quarterly*, *18*, 214–234.

Mariage, T. V. (2000). Constructing educational possibilities: A sociolinguistic examination of meaning-making in "sharing chair". *Learning Disability Quarterly*, *23*, 79–104.

Mariage, T. V. (2001). Features of an interactive writing discourse: Conversational involvement, conventional knowledge, and internalization in "morning message." *Journal of Learning Disabilities*, *34*, 172–196.

Marks, S. U., & Gersten, R. (1998). Engagement and disengagement between special and general educators: An application of Miles and Huberman's cross-case analysis. *Learning Disability Quarterly*, *21*, 34–56.

Mead, M. (1963). *Sex and temperament in three primitive societies*. New York: Morrow.

Medved, M. I., & Brockmeier, J. (2004). Making sense of traumatic experiences: Telling your life with Fragile X syndrome. *Qualitative Health Research*, *14*, 741–759.

Moody, S. W., Vaughn, S., Hughes, M. T., & Fisher, M. (2000). Reading instruction in the resource room: Set up for failure. *Exceptional Children*, *66*, 305–316.

Morningstar, M. E., Turnbull, A. P., & Turnbull, H. R. (1995). What do students with disabilities tell us about the importance of family involvement in the transition from school to adult life? *Exceptional Children*, *62*, 249–260.

Murdick, N., Shore, P., Chittooran, M. M., & Gartin, B. (2004). Cross-cultural comparison of the concept of "otherness" and its impact on persons with disabilities. *Education and Training in Developmental Disabilities*, *39*, 310–316.

Page, B., & Chadsey-Rusch, J. (1995). The community college experience for students with and without disabilities: A viable transition outcome? *Career Development for Exceptional Individuals*, *18*(2), 85–96.

Palincsar, A. S., Magnusson, S. J., Collins, K. M., & Cutter, J. (2001). Making science accessible to all: Results of a design experiment in inclusive classrooms. *Learning Disability Quarterly*, *24*, 15–32.

Palincsar, A. S., Parecki, A. D., & McPhail, J. C. (1995). Friendship and literacy through literature. *Journal of Learning Disabilities*, *28*, 503–510, 522.

Patching, B., & Watson, B. (1993). Living with children with an intellectual disability: Parents construct their reality. *International Journal of Disability, Development and Education*, *40*, 115–131.

Peck, C., & Furman, G. C. (1992). Qualitative research in special education: An evaluative review. In: R. Gaylord-Ross (Ed.), *Issues and research in special education*, (Vol. 2, pp. 1–42). New York: Teachers College Press.

Pugach, M. C., & Wesson, C. L. (1995). Teachers' and students' views of team teaching of general education and learning-disabled students in two fifth-grade classes. *The Elementary School Journal*, *95*, 279–295.

Reid, D. K., & Button, L. J. (1995). Anna's story: Narratives of personal experience about being labeled learning disabled. *Journal of Learning Disabilities*, *28*, 602–614.

Reiff, H. B., Gerber, P. J., & Ginsberg, R. (1993). Definitions of learning disabilities from adults with learning disabilities: The insiders' perspectives. *Learning Disability Quarterly*, *16*, 114–125.

Reiff, H. B., Gerber, P. J., & Ginsberg, R. (1994). Instructional strategies for long-term success. *Annals of Dyslexia*, *44*, 270–289.

Reiff, H. B., Ginsberg, R., & Gerber, P. J. (1995). New perspectives on teaching from successful adults with learning disabilities. *Remedial and Special Education*, *16*, 29–37.

Rice, D., & Zigmond, N. (2000). Co-teaching in secondary schools: Teachers reports and American classrooms. *Learning Disabilities Research & Practice*, *15*, 190–197.

Richardson, G. M., Kline, F. M., & Huber, T. (1996). Development of self-management in an individual with mental retardation: A qualitative case study. *The Journal of Special Education*, *30*, 278–304.

Ruiz, N. T. (1995a). The social construction of ability and disability: I. Profile types of Latino children identified as language learning disabled. *Journal of Learning Disabilities*, *28*, 476–490.

Ruiz, N. T. (1995b). The social construction of ability and disability: II. Optimal and at-risk lessons in a bilingual special education classroom. *Journal of Learning Disabilities*, *28*, 491–502.

Ruiz, N. T., Rueda, R., Figueroa, R. A., & Boothroyd, M. (1995). Bilingual special education teachers' shifting paradigms: Complex responses to educational reform. *Journal of Learning Disabilities*, *28*, 622–635.

Sabornie, E. J. (2004). Qualitative research and its contributions to the knowledge of emotional and behavioral disorders. In: R. B. Rutherford Jr., M. M. Quinn & S. R. Mathur (Eds), *Handbook of research in emotional and behavioral disorders* (pp. 567–581). New York: Guilford.

Sabornie, E. J., Cullinan, D., Osborne, S. S., & Brock, L. B. (2005). Intellectual, academic, and behavioral functioning of students with high-incidence disabilities: A cross-categorical meta-analysis. *Exceptional Children*, *72*, 47–63.

Sabornie, E. J., Evans, C., & Cullinan, D. (in press). Comparing characteristics of high incidence disability groups: A descriptive review. *Remedial and Special Education*.

Schwandt, T. A. (2001). *Dictionary of qualitative inquiry* (2nd ed.). Thousand Oaks, CA: Sage.

Scruggs, T. E., & Mastropieri, M. A. (1994a). Successful mainstreaming in elementary science class: A qualitative study of three reputational classes. *American Educational Research Journal*, *31*, 785–811.

Scruggs, T. E., & Mastropieri, M. A. (1994b). The construction of scientific knowledge by students with mild disabilities. *The Journal of Special Education*, *28*, 307–321.

Scruggs, T. E., & Mastropieri, M. A. (1995). Qualitative research methodology in the study of learning and behavioral disorders: An analysis of recent research. In: T. E. Scruggs & M. A. Mastropieri (Eds), *Advances in learning and behavioral disabilities* (Vol. 9, pp. 249–272). Greenwich, CT: JAI Press.

Shessel, I., & Reiff, H. B. (1999). Experiences of adults with learning disabilities: Positive and negative impacts and outcomes. *Learning Disability Quarterly*, *22*, 305–316.

Simpson, R. G., & Eaves, R. C. (1985). Do we need more qualitative research or more good research? A reaction to Stainback and Stainback. *Exceptional Children*, *51*, 325–329.

Skrtic, T. M. (1986). The crisis in special education knowledge: A perspective on perspective. *Focus on Exceptional Children*, *18*(7), 1–16.

Stainback, S., & Stainback, W. (1984). Broadening the research perspective in special education. *Exceptional Children*, *50*, 400–408.

Stainton, T., & Besser, H. (1998). The positive impact of children with an intellectual disability on the family. *Journal of Intellectual and Developmental Disability*, *23*(1), 56–69.

Stake, R. E. (1994). Case studies. In: N. Denzin & Y. Lincoln (Eds), *Handbook of qualitative research* (pp. 236–247). Thousand Oaks, CA: Sage.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Tedlock, B. (2003). Ethnography and ethnographic representation. In: N. Denzin & Y. Lincoln (Eds), *Handbook of qualitative research* (2nd ed., pp. 165–213). Thousand Oaks, CA: Sage.

Todis, B., Bullis, M., Waintrup, M., Schultz, R., & D'Ambrosio, R. (2001). Overcoming the odds: Qualitative examination of resilience among formerly incarcerated adolescents. *Exceptional Children*, *68*, 119–139.

Trout, A. L., Nordness, P. D., Pierce, C. D., & Epstein, M. H. (2003). Research on the academic status of children with emotional and behavioral disorders: A review of the literature from 1961 to 2000. *Journal of Emotional and Behavioral Disorders*, *11*, 198–210.

Turnbull, H. R., Guess, D., & Turnbull, A. (1988). Vox populi and Baby Doe. *Mental Retardation*, *26*, 127–132.

Ulman, J. D., & Rosenberg, M. S. (1986). Science and superstition in special education. *Exceptional Children*, *52*, 459–460.

Vaughn, S., Moody, S. W., & Schumm, J. S. (1998). Broken promises: Reading instruction in the resource room. *Exceptional Children*, *64*, 211–225.

Zetlin, A. G. (1986). Mentally retarded adults and their siblings. *American Journal of Mental Deficiency*, *91*, 217–225.

STATISTICAL ANALYSIS FOR SINGLE SUBJECT RESEARCH DESIGNS

Thomas E. Scruggs, Margo A. Mastropieri and Kelley S. Regan

ABSTRACT

Single subject research has long been employed to evaluate intervention effectiveness with students with learning or behavioral disabilities. Typically, the results of single subject research are presented on graphic displays and analyzed by a method of visual inspection, in which analysts simultaneously consider such data elements as level change, slope change, and variability in baseline and treatment data. However, over the years several concerns regarding visual inspection have emerged, including relatively low inter-rater reliabilities. This chapter reviews the arguments in favor of visual inspection as an analytic tool, and also summarizes the arguments favoring statistical analysis of single case data. The use of randomization tests is recommended, and an example is provided of its use in research with students with learning and behavioral disorders.

Throughout the history of special education research, single subject research methods have played an extensive and prominent role, in part because of its

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 33–53

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19002-3

particular relevance to individuals with learning or behavioral disabilities (Tawney & Gast, 1984). Special education, in contrast to other areas of education, is particularly concerned with the individual student, specific active interventions to improve learning and behavioral functioning, and the development of practical procedures that can be applied in a variety of real-world contexts (Lewis, Lewis-Palmer, Newcomer, & Stichter, 2004). Special education is also devoted to problem-solving, best addressed by ongoing research in applied settings (Horner et al., 2005). These areas are also of importance to the conceptual underpinnings of applied behavior analysis, as represented in single subject research.

Horner et al. (2005) listed ways in which single subject research provides an appropriate match with the needs of special education:

1. *Single subject research is typically oriented to the individual case.* This fact is significant for special education research, often concerned with low incidence disabilities, or low incidence behaviors (e.g., self-injurious behavior, self-stimulation, echolalia). Single subject research allows the individual to be treated as unit of analysis, with individualized treatments applied and evaluated.
2. *Single subject research can accommodate analysis of "nonresponders" as well as "responders."* When the behavior of groups is studied, individuals can respond very differently from the group, as is often the case in special education. Single subject research provides a methodology for studying those whose responses are not as predictable as those more representative of a given group.
3. *Single subject research provides a practical methodology for implementing and evaluating educational and behavioral treatments.* Students with learning and behavioral disabilities are in need of practical, effective, individually applied treatments, and single subject research methods are highly appropriate for these purposes.
4. *Single subject methodology allows for the evaluating experimental effects under standard educational conditions.* Using single subject methodology, special educators can evaluate the effects of recommended interventions over time in standard settings. Analysis of maintenance as well as initial effects is also possible.
5. *Single subject research designs can allow researchers to test conceptual theory.* Beyond the practical, applied application of behavioral principles, single subject research methodology can test the validity of behavioral theories that predict the conditions under which behavioral change would be expected in special education interventions.

6. *Single subject research methodology is a cost-effective method for identifying treatments that are appropriate for large-scale analysis*. Single subject research can be employed to develop a corpus of convincing evidence that justifies the application of larger-scale (and more expensive) designs relevant to special education.

Horner et al. (2005) also identified a number of "quality indicators" that should characterize credible single subject research. These involved indicators in the areas of (a) description of participants, (b) description and measurement of dependent variables, (c) systematic description and implementation of the independent variables, (c) appropriate description and implementation of baseline phases, (d) demonstration of experimental control, (e) demonstration of external validity, and (f) social validity. With respect to experimental control, however, Horner et al. (2005) suggest that the data demonstrate experimental control, that threats to internal validity are addressed, and that at least three demonstrations of experimental effect are provided, at three different points in time. The suggested method for evaluating experimental effects is the "traditional approach" of visual analysis.

CASE FOR VISUAL ANALYSIS OF SINGLE SUBJECT DATA

Most single subject research is evaluated by means of visual analysis procedures, or the visual consideration of data presented in a graphic format. A strong case for the overwhelming reliance of single subject researchers on visual inspection methods was made by Busk and Marascuilo (1992), who reviewed all articles published in the *Journal of Applied Behavior Analysis* during 1988. All articles that involved single subject methodology employed visual inspection of the data as the method of analysis, although statistical analysis was employed to evaluate between-group data in the same journal. Busk and Marascuilo (1992) also referred to a study by Kratochwill and Brody (1978), who evaluated single subject research published in four behaviorally oriented research journals, *Behavior Therapy, Behaviour Research and Therapy, Journal of Applied Behavior Analysis,* and *Journal of Behavior Therapy and Experimental Psychiatry.* They reported that single subject research included statistical analysis in only a small minority of cases in each journal, between only 4 and 9%. A review of the most recent volume of *Behavior Modification* supports the conclusion that visual inspection methods remain the predominant method of analysis for single subject data.

General guidelines for using visual analysis, or visual inspection, have been provided by a number of previous authors, including Hersen and Barlow (1984), Parsonson and Baer (1978, 1986, 1992), Tawney and Gast (1984), Horner et al. (2005), and Kennedy (2005), to name only a few. Typically, researchers are encouraged to present the obtained data in a graphic, time-series format, in which the behavior being observed is presented on the ordinate (or vertical axis) and time is presented on the abscissa (or horizontal axis) of the data display. Researchers then examine the data visually and simultaneously evaluate multiple components of the graphic presentation of data, considering particularly within-phase trends and variability, across-phase level and slope changes, and across-phase data overlap. Taking all these features into account, then, the researcher makes a determination regarding the degree of effectiveness of the intervention.

A good deal of agreement exists on the features of data to be considered when conducting a visual analysis. Parsonson and Baer (1978) combined the suggestions of a number of different researchers and concluded the following features were relevant:

1. *Stability of baseline.* Baseline data should remain stable (i.e., employ minimal variability and no obvious trend) or trend in the direction opposite anticipated treatment effects.
2. *Within-phase variability.* Highly variable baselines may suggest controlling variables may have been in effect. Baselines that stabilize after initial variability may suggest a transitory effect of the observation process. Treatment data that are highly variable may imply the need for additional examination or treatment to stabilize the behavior. Behavior that begins as stable and then becomes more variable may suggest that satiation of the reinforcer or inhibition are influencing behavior.
3. *Between-phase variability.* Experimental control is apparent when stable treatment effects follow variable baseline data. Evidence of control is less apparent when baseline and treatment data are both variable.
4. *Across-phase data overlap.* The amount of overlapping data between baseline and treatment phases is a prime indicator of intervention effectiveness. Parsonson and Baer (1978) noted, "although there are no established criteria of excessive or acceptable amounts of overlap, the less overlap, the more convincing the treatment effect" (p. 122, see also Scruggs & Mastropieri, 2001).
5. *Number of data points.* More data points in each phase contribute to more valid interpretations of the data. Generally, more data points are

necessary when data are variable, there is substantial overlap, or the behavioral data appear to "drift."

6. *Within-phase change in trend.* Within phase trends can suggest inappropriate baseline drift in the direction of treatment effect, making interpretation difficult, or trends in the intervention phase, suggesting additional sources of behavioral control.
7. *Changes across adjacent phases.* A sharp change in trend between baseline and treatment phases can indicate substantial behavioral control. However, in other cases, a curvilinear function could be apparent, suggesting little behavioral control.

Single subject researchers typically argue that such visual inspection methods are most sensitive to large, easily observable treatment effects, of the degree of magnitude where one could expect little disagreement. It is argued that these measures are conservative, and as such, reduce the probability of "Type I" errors, that is, that ineffectual treatments will be interpreted as effective. For example, Parsonson and Baer (1978) suggested that visual analysis "usually is relatively insensitive, yet that very lack of refinement may have important and valuable consequences for the analysis of behavior" (p. 111). Furthermore, by striving for large and obvious treatment effects, researchers can be more confident that they are conducting research of some social significance. These arguments can hardly fail to appeal to single subject researchers investigating ways of improving levels of functioning of individuals with learning and behavioral disabilities.

Another stated advantage of visual inspection procedures is that they allow for a more precise and intimate interpretation of the data. According to Parsonson and Baer (1986), by employing visual analysis, "the audience is free … to make different interpretations, look into the fine grain of the data, and perhaps start down a new line of research or balk at what now seems to them an overinterpreted and even thus eventually useless line" (p. 165). The opportunity for such "fine grained" analysis is often reported as an advantage of visual inspection procedures (Parsonson & Baer, 1992).

A final argument favoring visual inspection of graphed data is that, over the decades following the advent of applied behavior analysis, these analytical procedures have been used to develop and validate the myriad of behavioral techniques and principles that are now commonly employed in schools, clinics, and other relevant settings. According to Kennedy (2005, p. 12):

> … research using single-case designs has provided tremendous insights into processes that improve educational practices and outcomes for a wide variety of students. For

> decades, this approach to experimental design has yielded easier-to-implement and more effective interventions, a deeper understanding of behavioral processes, more accurate and usable measurement systems, and greater benefits for students, families, and schools.

Thus it could be argued that the actual impact on practice is sufficient justification for methodology, and its analytical or evaluative components.

CONCERNS RAISED ABOUT VISUAL ANALYSIS

Unfortunately, the application of visual analysis procedures to graphic data displays has raised some concerns. Although the application of a number of features of visual analysis has been widely described, clear criteria for decision making for any of these criteria has been lacking. That is, while it can generally be agreed that such features as slope, level, variability, number of data points, and data overlap may be relevant, how these standards are to be employed is subject to question. How great of a level change? How much variability is of importance? When are slopes interpretable, and when are they not? How many data points are necessary for clear interpretation, and under what circumstances? For one example, consider the commonly heard argument that large effects are desirable. While this argument seems logical on the surface, what precisely constitutes a ‘‘large’’ effect? Yeaton (1982) argued, ‘‘Applied behavior analysts have not defined what they mean by a big effect’’ (p. 86). Indeed, a search of the methodological literature (e.g., Parsonson & Baer, 1978; Kennedy, 2005) reveals a lack of criteria with which a researcher can objectively conclude that a particular effect was ‘‘big.’’

Perhaps in part because of limited evaluative criteria, interrater reliability for treatment outcomes can be a concern. According to Kazdin (1978), ‘‘the problem with visual inspection is that individuals who peruse the data may not see eye to eye’’ (p. 638). Gottman and Glass (1978) reasoned similarly: ‘‘Clearly, the ‘eyeball test’ gives results that vary from judge to judge and that can conflict sharply with statistical tests’’ (p. 199). When studied empirically, interrater agreement for study outcomes has been found to be discouragingly low (DeProspero & Cohen, 1979; Franklin, Gorman, Beasley, & Allison, 1996; Jones, Vaught, & Weinrott, 1977; Jones, Weinrott, & Vaught, 1978; Yeaton, 1982), even when interrater reliability of behavioral observations is high. This appears to be a problem of particular significance, since a methodology which lacks consistently applied or understood evaluative criteria would appear to be limited in its contribution to scientific knowledge. Kennedy (2005), however, argued that in these studies, conclusions were drawn out of the context of the experimental process, and that the disagreements were most pronounced when the effects were subtle and there was high variability within phase. These are instances when most behavior analysts would unlikely to maintain that a functional relationship had been demonstrated.

Nevertheless, some single subject researchers are not entirely troubled by low reliability of visual analysis. Parsonson and Baer (1986, p. 165) stated, "we do not wish to agree that much: we value some of our differences in standards; we each value our freedom to believe or disbelieve a given interpretation of a set of data." Although there may be some merit in such a consideration, it remains true that some data, even when randomly generated, appear to represent a meaningful pattern that in reality has no meaning. Although it can be argued that individuals may reasonably take different views of the merits of reliable outcomes (this frequently happens in group-experimental research), surely it must be agreed that means are necessary to separate meaningful from nonmeaningful data patterns.

Todman and Dugard (2001) illustrated this issue in a dramatic fashion. Consider Parsonson and Baer's (1978) first feature of visual analysis, stability of baseline, or a steady state of behavior typically exhibited before intervention is implemented. This principle of behavior analysis was developed in previous laboratory research (see Sidman, 1960), and suggested that behavior should occur in a steady state, or highly predictable pattern, against which any pattern of responding can be indexed when the intervention is implemented. Although this standard may not be realized ethically in some situations (e.g., self-injurious behavior), the logic of baseline stability prior to intervention as an aid to interpretation has been accepted generally as an important criterion of behavior analysis (Kennedy, 2005).

However, Todman and Dugard (2001) argued that the practice intervening after a pattern of stability had been established could result in unintended interpretive consequences, in that, in random data, a pattern of little variability is frequently followed by an increase in variability. To test this hypothesis, they randomly generated 80 single subject data displays based upon an assumption of an intercept of 5 and a slope of 0.20. Each data point was allowed to vary randomly, either positively or negatively, in accordance with a normal curve with a mean of zero and a standard deviation of 1.5 (random normal deviate).

Of the 80 randomly generated graphs, 50 revealed several consecutive data points of little variability that could be used, by the Parsonson and Baer (1978) criteria, to designate a point for intervention in an A-B (baseline-treatment) design. Two copies were then presented of these 50 data displays: one indicated intervention lines placed at the end of the low variability (flat) points; the other indicated intervention lines placed at random. Twelve graduate students familiar with evaluating similar data displays were then asked to examine each graph and place it into one of two piles: those indicating a treatment effect, and those indicating no treatment effect. The results were provocative: 33 of the 50 (66%) of the graphs with the intervention line placed after the flat data points were sorted into the "effective" pile; only 11.4 (22.8%) of the data displays of randomly generated intervention lines were sorted into the effective pile.

It can (and perhaps should) be pointed out that, although randomly generated data may sometimes appear to exhibit an effect using the little-recommended A-B design, that such effects would be very unlikely to be manifest were a reversal (A-B-A-B) design employed. However, with respect to this argument, two points should be considered: (a) Todman and Dugard established that randomly generated data can frequently appear to be meaningful and systematic, and (b) the A-B design, although rarely recommended by single subject researchers, nevertheless is commonly found to represent individual participants (or behaviors) in multiple baseline designs. A third point, that perhaps waiting for baseline stability may lead to pernicious results, will be addressed later in this chapter.

A possible solution to this dilemma is to employ some type of systematic statistical analysis of single case data, in order to discriminate between reliable and nonreliable data patterns. This suggestion, however, is in itself controversial, and strong arguments have been raised on both sides (Kazdin, 1984).

THE CASE AGAINST STATISTICAL ANALYSIS

Many single subject researchers have argued against the use of statistical procedures (Kazdin, 1984; Michael, 1974); particularly influential was Skinner's (1963) statement that "statistical methods are unnecessary" (p. 508). Kazdin (1984) summarized the arguments against using statistics to analyze single subject data. Researchers, Kazdin suggested, are concerned with clinical or experimental significance. Researchers are interested in facilitating behavioral change that attains a socially desirable criterion. That is, individuals in a subject's everyday life (including parents, peers, friends, classmates, or teachers) determine in part which behaviors can be considered offensive, aberrant, laudatory, obnoxious, or acceptable. These individuals also generally set criteria for which degree of behavior constitutes a criterion of acceptability. The role of the researcher using applied behavior analysis is to bring about behavior change to the level at which relevant individuals view the behavior as no longer a problem. Typically, this requires substantial, obvious behavior change against relatively clear social standards. Weighed against these criteria, statistical methods seems unnecessary or even irrelevant. That is, if an individual who exhibited undesirable or even dangerous behavior now exhibits socially acceptable behavior, the application of statistical tests to these data can be viewed as redundant and inappropriate. Even for experimental purposes, Kazdin (1984) argued that the standard for behaviorists for obvious, replicable outcomes argues strongly against the necessity of statistical methods.

Finally, Kennedy (2005) raised a practical issue:

> The practical problem with using inferential statistics in single-case designs is that the currently existing statistics either violate fundamental statistical assumptions or are intractable in the large majority of applied research … . Therefore, the use of inferential statistics in single-case designs is largely an academic debate and not a practical issue for researchers looking for new analytical tools. (p. 192)

According to this argument, then, the case for statistical analysis is moot because practical statistical methods for evaluating single subject data are not available.

THE CASE FOR STATISTICAL ANALYSIS OF SINGLE SUBJECT DATA

Nevertheless, Kazdin (1984) and others have acknowledged that there may be circumstances in which statistical methods may be helpful. These circumstances involve situations in which behavior change is more subtle, such as in the initial stages of investigating a new treatment method, or in real-world situations where complete experimental control is lacking. In addition, statistical analyses to clarify ambiguous data may be helpful in cases where the design was altered for ethical reasons; for example, if baselines were not extended as long as they should have been for scientific purposes, because of a need to intervene on troubling behavior.

Huitema (1986b) made an argument favoring statistical analysis for all single subject data. Arguing that applied behavior analysis was to a large degree isolated, the use of statistics may provide more ready with others (e.g., professionals, researchers, scholars, practitioners, and funding agencies) who do value the application of statistical methods. Since statistics do not necessarily detract from the research investigation, it could provide a function that would be either neutral or augmentative:

> If an audience is convinced that a researcher has something worthwhile to say on the basis of statistical significance alone, consider how impressed it will be when both statistical significance and a very large visually apparent effect is presented. (p. 229)

PROPOSED STATISTICAL ANALYSES

To a certain extent, Kennedy (2005) is correct that there are not presently statistical procedures that all will agree could be applied appropriately to the analysis of single subject data. Nevertheless, there are a number of statistical procedures that should be considered, and at least one type of statistical analysis (randomization tests) that would appear to have the potential for broad application.

Analysis of Variance and Autocorrelation

Among the earliest advocates of statistical methods for single subject research were Gentile, Roden, and Klein (1974), who recommended the use of traditional inferential statistics, such as *t* and *F* tests, to test between-phase performance differences. This method was generally criticized on the basis of the assumption that single subject data, which necessarily employ multiple measures of behavior of individual subjects, is necessarily non-independent, a necessary assumption of analysis of variance (e.g., Gorman & Allison, 1996). Gentile et al. (1974) maintained that serial dependence could be reduced by combining data across phases, arguing that data on individual participants collected across time intervals are less strongly related than data that are immediately adjacent. This argument, of course, does not address the issue of possible autocorrelation within phases.

However, arguments have been raised that single subject data are not always, or even typically, autocorrelated. Degree of autocorrelation can be directly calculated, typically by computing correlation coefficients between each data point and the point following it. If each data point is correlated with each subsequent data point, the result is referred to as a *lag one* autocorrelation. If each data point *n* is correlated with the second data point following ($n+2$), the result is referred to as a *lag two* autocorrelation, and so on. Typically, the smaller the lag, the higher the autocorrelation in serially dependent data, with the possible exception of "seasonal effects" (e.g., weekly or monthly effects). Therefore, serially dependent data would be expected to display higher lag one autocorrelations, and lower correlations as the lag numbers increase (Kazdin, 1984). Huitema (1985, 1986a) analyzed

441 single subject data displays including 1748 different experimental phases takes from 10 volumes of the *Journal of Applied Behavior Analysis*. He constructed a histogram of within-phase, lag one autocorrelations of the observed data, and concluded that the histogram revealed a near-normal distribution, with a mean near zero. From this he concluded that single subject data are typically uncorrelated, and, presumably, standard statistical analyses are appropriate.

Busk and Marascuilo (1992), however, challenged the contention that single subject data are not typically correlated. First, they suggested, there may be simply too few data points to establish serial dependency. Even if a trend is evidenced, the small number of data presented may preclude statistical significance; however, a non-significant correlation is not necessarily the same as a zero correlation. The duration of the interval between the observations may also influence autocorrelations. Further, revealing that autocorrelations reveal a mean near zero may not indicate that autocorrelations are not found; it may mean, rather, that there are similar numbers of positive and negative autocorrelations, with a mean near zero. Finally, Busk and Marascuilo (1992) reported on their own (1988) analysis of 248 data sets from 44 research studies. They computed lag one autocorrelations and reported that 80% of the autocorrelations ranged between 0.10 and 0.49 for phases containing between 5 and 18 observations. In addition, 40% of the baseline data phases revealed autocorrelations larger than 0.25; 59% of the intervention data phases were greater than 0.25. Busk and Marascuilo (1988, 1992) concluded that statistical tests assuming independence performed on these data would result in an inflated Type I error (see also Matyas & Greenwood, 1996).

Other Statistical Tests

Apart from traditional analysis of variance, a number of other statistical tests have been proposed for analyzing single subject data. One approach is the autoregressive, integrated moving average (ARIMA) model (Box & Jenkins, 1976). This is essentially a regression approach that characterizes that an observation at a given point in time is a function of the values of the previous observations and accumulated residual errors. This approach involves building equations with autoregressive terms, differencing values (removing autoregressive components), and modeling residuals (Gorman & Allison, 1996). Although ARIMA models can address the issue of serial dependency, a very substantial number of data points in baseline and intervention phases are required, and would not be practicable for most single subject research designs. The examples of these procedures provided by Kazdin (1984) and Gorman and Allison (1996) each displayed 20 data points in both baseline and treatment conditions. In contrast, Huitema (1985, 1986a) reported that over half of the baseline phases he evaluated contained fewer than 6 observations, a figure far too small for regressive approaches such as ARIMA (see also Scruggs, 1992).

Further, in many single subject investigations, it is difficult to collect so many baseline observations in cases where there is an urgent need for behavioral change. Although such approaches may be valuable for a small number of planned investigations, they seem unlikely to have broad applicability. Gorman and Allison (1996) described a number of statistical alternatives, and concluded that the major challenges to statistical analysis of single subject data include sample size (number of observations), the need for randomization of treatments, and autocorrelation. Although some procedures addressed these concerns better than others, Gorman and Allison concluded, “all techniques are severely limited by small sample sizes” (1996, p. 205), and recommended that single subject researchers begin to collect additional data. Busk and Marascuilo (1992) suggested “at least 35 to 40 observations are needed for each phase in order to justify the time-series model” (p. 166), a condition unlikely to be met in single subject research. One possible alternative, however, that can potentially address these challenges, may be the use of randomization tests.

RANDOMIZATION TESTS

Randomization tests have been proposed for many years (Edgington, 1992); however, more appropriate applications and the capacity for computation have appeared more recently. In some cases, proposed models (e.g., 12 permutations of A-B phases with one data point per phase) seemed unlikely to be applied in behavioral research. In others, the demands on computation seemed excessively great. One example provided by Levin, Marascuilo, and Hubert (1978) admittedly addressed an alpha level of only 0.33. Kazdin (1984), described, among other examples, a case where 4 treatment and 4 baseline conditions were implemented at random over 8 days, a design unlikely to be employed in much single subject research. However, other models have shown great promise as a supplement to traditional visual inspection of data, without stringent requirements that alter the nature of the experiment.

A-B designs. Edgington (1992) provided a simple example of a randomization test for an A-B design. Consider a single subject investigation consisting of 25 daily observations. In order to allow a sufficient number of baseline or treatment observations for interpretation of the data, the researcher elects to choose a point of intervention between the 6th and the 20th data point, assuring there will be at least 5 baseline and 5 intervention observations. This intervention point is then chosen at random from the 20 possible points. Suppose that the point selected was at day 9, with 8 baseline data points and 17 treatment data points, and the results are as follows (underlined values represent treatment observations):

Day:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Value:	4	3	3	2	4	4	3	3	8	7	7	8	9	6	7	7	8	8	7	9	8	7	8	12	9

Mean values for treatment and baseline can be computed, at 7.941 for treatment mean and 3.250 for baseline mean, with a mean difference of 4.691. This sum can then be compared with all other possible mean differences when intervention points between 6 and 20 are chosen. For example, for intervention at day 6, the treatment mean would be 7.250 and the baseline mean would be 3.200, with a difference of 4.050, smaller than the observed value of 4.691. For intervention at day 7, the treatment mean would be 7.421 and the baseline mean would be 3.333, with a difference of 4.088, also smaller than the observed value. When all 20 permutations are computed, it would be seen that the observed value was the highest of all phase mean differences, and since the intervention point was selected at random from 20 possible intervention points, the associated probability of this value being a chance occurrence is 1/20, or 0.05, in a one-tailed test (for a two-tailed test, the absolute value of the phase mean differences can be computed).

This example is intentionally simple for demonstration purposes; in fact, few single subject researchers would employ an A-B design for only one subject as a scientific investigation. Also, in this case, the power to detect a significant difference is relatively low. Applications to other, more complicated designs have also been described.

A-B-A designs. Todman and Dugard (2001) described the use of randomization tests for a number of experimental designs. In the case of an A-B-A design, the researcher may wish to establish baseline function, then implement an intervention, then return to baseline conditions to determine whether removal of the intervention contingency results in a return to baseline levels. Using procedures similar to the preceding example, the researcher selects intervention and withdrawal points at random, within established parameters. For example, for a proposed investigation of 36 data points, for which it is stipulated that at least 8 data points will be included in each of the three treatments, Todman and Dudgard (2001) demonstrated that there were 91 possible pairings of intervention and withdrawal initiation points. Then, the mean of the intervention phase is computed, subtracted from the combined preintervention–postwithdrawal phases, and this difference is compared with the same difference derived from all possible permutations of intervention and return to baseline initiation points. For a one-tailed test, all randomly attained values are compared with the observed value to determine how many were higher (or lower) than the observed value. For a two-tailed, differences in absolute value are computed. In this instance, the lowest attainable alpha level is 0.011, in the case that no other randomly generated value exceeded the observed value.

Multiple baseline designs. A design with a high potential for practical application described by Todman and Dugard (2001) is the randomization test for A-B multiple baseline designs (see also Edgington, 1992; Marascuilo & Busk, 1988). They demonstrated that with 3 baselines, only three possible intervention points for each participant would be necessary for some measure of statistical power. That is, if the three intervention points were selected at random, the total number of possible arrangements would be 33, or 27 possible arrangements, with a possible *p* value as low as 1/27, or 0.037, in a randomization test, given the observed value was the greatest. With higher numbers of participants, and larger numbers of possible intervention points, the number of possible combinations can become so great that individual calculation of all values is no longer practicable. Using software included with the Todman and Dugard (2001) book to be incorporated with other software for statistical analysis, such as SPSS for Windows, as many as 2,000 random permutations of treatment-baseline mean differences can be generated and compared to the actual observed value. An example of this method is demonstrated in a following section.

Issues in using randomization tests. In spite of the potential advantages of randomization tests, some concerns have been raised. Gorman and Allison (1996) presented some data that suggested that autocorrelated data may affect the Type I error rate in some cases. Levin et al. (1978), however, concluded autocorrelation of individual observations was not a problem where data are aggregated within phases, and Busk and Marascuilo (1992) concurred that randomization tests conducted on phase means are generally appropriate. Todman and Dugard (2001) concluded that the conclusions of Gorman and Allison (1996) were arguable since there was no randomization within phases, and the individual observation constituted the unit of analysis. Further research may provide additional information on this issue, but at present it appears that randomization tests are appropriate for single subject data (Busk & Marascuilo, 1992).

Some other issues are worthy of consideration. Although there are probably many more opportunities for random assignment of intervention points than has been argued, it is certain that there are times when ethical considerations preclude this. For example, if an individual participant is in danger of injuring him/herself or others (e.g., severe self-injurious behavior), intervention must be undertaken as soon as possible. Although the treatment element of the research is clearly of most importance, it is also true that the design and interpretation of the study may be more limited. In such cases, perhaps it would be appropriate to choose random intervention points in cases which are similar but when potential for harm is lessened, and then compare these results with those of interventions with less internal validity and more urgency, which nevertheless employed similar treatments. Nevertheless, the data presented by Todman and Dugard (2001) suggest strongly that the previous insistence of single subject researchers upon stability of baseline observations prior to implementation of intervention, be reconsidered.

There also may be some cases for which randomization tests are less appropriate. For example, there may be instances where there appears to be a need to evaluate an important treatment, but not enough time, resources, or participants to design and evaluate outcomes statistically. In such cases, considerations for statistical analysis should not preclude implementing appropriate interventions when needed, and generating and providing such evaluation data as are possible under the circumstances.

In other instances, randomization tests as described here simply may not be entirely appropriate, or may require some modification. Interpretation may be difficult when, for example, baseline data decline sharply, followed by a similarly sharp increase under intervention. If the intervention was in fact the reason for the sharp change in trend, the associated mean difference may be zero, with randomly selected intervention points revealing similar numbers of positive and negative values. In cases where within-phase trends are apparent (or, optimally, can be predicted), Levin et al. (1978) suggested that slope, rather than mean, would be the appropriate metric. In another instance, it may be that initiation of treatment does not result in an immediate change in data, but rather a delayed effect, 3 or 4 observations after treatment is initiated. In such cases, randomization tests may not reveal appropriate probabilities. However, it is true that randomization tests are flexible and can be designed to meet a number of contingencies (Todman &

Dugard, 2001). In this instance (particularly if a delayed effect of treatment could be predicted), it may be possible to compare means before and after the 4th observation following treatment initiation, and compare this observed difference with the mean value before and after the 4th observation after a number of randomly generated initiation points.

APPLICATIONS OF THE RANDOMIZATION TEST

Regan, Mastropieri, and Scruggs (2005) recently implemented a single subject research investigation employing randomization tests as a component of the data analysis. This study investigated the effects of dialogue journals on expressive writing of five students with behavioral disorders. These students included one female and four males, identified as having emotional or behavioral disorders (EBD), between the ages of 11 and 12, attending a self-contained sixth-grade class for students with EBD, and had behavioral goals included on their individualized education plans (IEPs). These behaviors included immature and attention-seeking behavior, problems with maintaining boundaries and social interactions, excessive anger, problems remaining on task, and using coping strategies when frustrated. For all five of these students, it was anticipated that the use of dialogue journals would facilitate expressive writing, promote work on target behaviors through the communicative exchange, and improve communication with the teacher. Importantly for the purposes of randomization tests, intervention points were selected at random for each student.

Under baseline, traditional writing prompts were provided in each student's journal. This consisted of a request or question to which students were required to provide a written response (e.g., "Write about an interesting time in your childhood," What would be in your perfect world and why?").

During intervention, dialogue journals were implemented. A dialogue journal is an ongoing written conversation between students and teachers, with an emphasis on personal content relevant to participants. In this case, the content of the dialogue journals focused on the social/behavioral issues identified on student IEPs and by teacher observations. Specific written guidelines included the following:

1. The journal, a 70-page spiral notebook, is a private writing between the teacher and you, the student.
2. We will communicate four days a week at the set journaling time of 15 minutes.

3. Each of us will write back and forth with a minimum of five sentences, no maximum.
4. We can respond to questions and comments, introduce new topics, and/or ask questions.
5. My focus will be on the quality of the writing rather than punctuation, spelling, and handwriting. Responses will not be graded.
6. We will both write to each other in a 'letter format' with the date, a greeting (ex., "Hi, Ms. Smith,"), a closing (ex., "Sincerely,"), and signature/sign-off.
7. If any information comes up that puts you or others at risk, this information may need to be shared with other adults.
8. The main goal for the dialogue journal will be meaningful communication. (Regan et al., 2005, p. 38).

The intervention was implemented for approximately six weeks. Dependent variables included time on task during journaling time, length of student writing in number of words written, and quality of student writing as rated by trained graduate students using a scoring rubric. For quality of student writing, each student entry was typed over, and then corrected for mechanics, including spelling, punctuation, and capitalization, so that handwriting quality and writing mechanics did not influence overall quality rating. The quality rating was based upon a judgment of the overall quality of the writing, including word choice, grammar, sentence structure, organization, and ideation. Generalization data were also collected during the writing workshop of a language arts class which students took in a different classroom. Reliability of observations of time on task behavior was assessed at 93%, and reliability of scoring of writing samples.

Both visual analyses and randomization tests were employed for data analyses across the dependent variables. Visual analyses employed criteria described previously, including evaluation of changes across phases with respect to slope, level, and variability, considered simultaneously. The randomization tests were conducted to determine the probability that any other randomly selected arrangement of intervention points would have resulted in greater baseline-treatment differences. In the present instance, the randomization tests addressed performance of students employed in a multiple baseline design. Mean differences in performance between treatment and baseline phases were computed and summed across participants. In other words, the average of baseline data (for example, time on task) is subtracted from the average of intervention data. These differences are summed across participants to calculate a total sum. Since the time for initiation of intervention was selected at random for each participant, this sum can be compared with all other random permutations of intervention points between the first (5) and final (21) intervention point. However, since over 200,000 permutations are possible (an unnecessarily high number), the program developed by Todman and Dugard (2001), using SPSS for Windows (SPSS, 2005) randomly selects 2,000 intervention points, and compares the number of times the total sums of behavior differences from randomly selected intervention points exceeds the observed value. This total sum, then, is compared with the sums calculated from 2,000 randomly selected arrangements of intervention points and calculates the number of times randomly generated data exceed the obtained value. In the case of the Regan et al. (2005) data, values higher than those observed were not generated with respect to attention to task, number of words written, nor writing quality, with all p values < 0.001. Specifically, the exact probability of the observed experimental data occurring by chance was 1/2,001 (i.e., 2,000 randomly generated summed differences + 1 actual summed difference) = 0.00049975, in each case.

As can be seen, the results were highly statistically significant. It was not necessary for every randomly generated value to exceed the observed value; in this case it was only necessary for no more than 100 of the 2,000 randomly generated values to exceed the observed value for the result to be statistically significant according to a conventional standard ($p<0.\ 05$). Visual inspection procedures, applied to the same data, also led to the conclusion that the intervention had had an impact on the dependent variables. However, the obtained data when considered visually, did not demonstrate very obvious, pronounced effects, as may have been expected in the present academic intervention, where a specific type of written dialoging was being compared with a similar, albeit less personal, activity. The use of visual inspection techniques, coupled with the strong statistical outcomes, argued strongly that the intervention had been effective. The randomization test, in this case, confirmed the impact of the treatment, without intruding on the ability of the researcher (or readers) to interpret the value or meaningfulness of the treatment, in terms of magnitude of effect compared with social or academic criteria.

SUMMARY

In the analysis of single subject research, there is no substitute for large, replicable, visually apparent effects, that establish without question the functional relationship between dependent and independent variables. In such cases, statistical analysis can be considered superfluous, redundant, or unnecessary. Additionally, clinical circumstances and ethical concerns may require that treatment be initiated and evaluated in a specific way, which preclude the use of statistics. In other cases, however, where there is opportunity to plan a more carefully designed intervention, and where large, obvious treatment effects can not be predicted with certainty, some type of statistical analysis, particularly using randomization tests, may be very appropriate and justifiable. Other methods of statistical analysis of single subject data presently appear inappropriate (e.g., analysis of variance), or require data characteristics typically not present in most single subject research (e.g., time series analysis).

Data provided by Todman and Dugard (2001) provide evidence for concern about the traditional practice of waiting until baseline data stabilize before implementing intervention. Given these data, it seems possible or even likely that at least some previous insistence on stable baselines may have resulted in the spurious appearance of intervention effects. Random assignment of treatment initiation times may relieve this concern and allow for the use of statistical tests that can determine that observed effects are statistically reliable. Given demonstration of reliable treatment effects, then, researchers are free to interpret the practical or social importance of the observed data and the implications for further research.

Statistical analysis of single subject data is maximally effective when it is used as a supplement to, rather than a replacement for, visual analysis. In these cases, both procedures can complement each other. Visual analysis can continue to focus the attention of researchers on behavior change of social importance; statistical analysis can provide a check on the reliability of obtained data. Further, statistical analysis can assist with the translation of research to other researchers or consumers of research. The recent emphasis, for example, of the U.S. Department of Education on random assignment (e.g., Whitehurst, 2005), suggests that statistical analysis based upon randomization tests may result in single subject research evidence of more influence and impact than single subject data that involves no randomization and that is evaluated entirely through visual inspection. Future applications of appropriate statistical analyses of single subject data may demonstrate the practical utility of these techniques to identify reliable treatments, and to increase the influence of single subject research.

REFERENCES

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis, forecasting and control.* San Francisco: Holden-Day.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, *10*, 229–242.

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In: T. R. Kratochwill & J. R. Levin (Eds), *Single-case research design and analysis: New directions for psychology and education* (pp. 159–186). Mahwah, NJ: Lawrence Erlbaum Associates.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, *12*, 563–570.

Edgington, E. S. (1992). Nonparametric tests for single-case experiments. In: T. R. Kratochwill & J. R. Levin (Eds), *Single-case research design and analysis: New directions for psychology and education* (pp. 133–158). Mahwah, NJ: Lawrence Erlbaum Associates.

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In: R. D. Franklin, D. B. Allison & B. S. Gorman (Eds), *Design and analysis of single-case research* (pp. 119–158). Mahwah, NJ: Lawrence Erlbaum Associates.

Gentile, J. R., Roden, A. H., & Klein, R. D. (1974). An analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, *5*, 193–198.

Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single-case designs. In: R. D. Franklin, D. B. Allison & B. S. Gorman (Eds), *Design and analysis of single-case research* (pp. 159–214). Mahwah, NJ: Lawrence Erlbaum Associates.

Gottman, J. M., & Glass, G. V. (1978). Analysis of time-series experiments. In: T. R. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 197–235). New York: Academic Press.

Hersen, M., & Barlow, D. H. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–179.

Huitema, B. E. (1985). Autocorrelation in behavioral research: A myth. *Behavioral Assessment*, *7*, 109–120.

Huitema, B. E. (1986a). Autocorrelation in behavioral research: Wherefore are thou? In: A. Poling & R. W. Fuqua (Eds), *Research methods in applied behavioral analysis: Issues and advances* (pp. 187–208). New York: Plenum.

Huitema, B. E. (1986b). Statistical analysis and single-subject designs: Some misunderstandings. In: A. Poling & R. W. Fuqua (Eds), *Research methods in applied behavioral analysis: Issues and advances* (pp. 209–232). New York: Plenum.

Jones, R. R., Vaught, R. S., & Weinrott, M. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, *10*, 151–166.

Jones, R. R., Weinrott, M., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, *11*, 272–282.

Kazdin, A. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology*, *46*, 629–642.

Kazdin, A. (1984). Statistical analysis for single-case experimental designs. In: M. Hersen & D. H. Barlow (Eds), *Single case experimental designs: Strategies for studying behavior change* (2nd ed., pp. 265–317). New York: Pergamon Press.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston: Allyn & Bacon.

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification*, *2*, 291–307.

Levin, J. R., Marascuilo, L. A., & Hubert, L. J. (1978). N = Nonparametric randomization tests. In: T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change* (pp. 167–196). Orlando, FL: Academic Press.

Lewis, T. J., Lewis-Palmer, T., Newcomer, L., & Stichter, J. (2004). Applied behavior analysis and the education and treatment of students with emotional and behavioral disorders. In: R. B. Rutherford Jr., M. M. Quinn & S. R. Mathur (Eds), *Handbook of research in emotional and behavioral disorders* (pp. 523–545). New York: Guilford.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, *10*, 1–28.

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In: R. D. Franklin, D. B. Allison & B. S. Gorman (Eds), *Design and analysis of single-case research* (pp. 215–243). Mahwah, NJ: Lawrence Erlbaum Associates.

Michael, J. (1974). Statistical inference for individual organism research: Some reactions to a suggestion by Gentile, Roden, and Klein. *Journal of Applied Behavior Analysis*, *7*, 629–634.

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In: T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change* (pp. 101–166). Orlando, FL: Academic Press.

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In: A. Poling & R. W. Fuqua (Eds), *Research methods in applied behavior analysis: Issues and advances* (pp. 157–186). New York: Plenum.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In: T. R. Kratochwill & J. R. Levin (Eds), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Mahwah, NJ: Lawrence Erlbaum Associates.

Regan, K. S., Mastropieri, M. A., & Scruggs, T. E. (2005). Promoting expressive writing among students with emotional and behavioral disturbance via dialogue journals. *Behavioral Disorders*, *31*, 35–52.

Scruggs, T. E. (1992). Single subject methodology in the study of learning and behavioral disorders: Design, analysis, and synthesis. In: T. E. Scruggs & M. A. Mastropieri (Eds), *Advances in learning and behavioral disabilities* (Vol. 7, pp. 223–248). Oxford, UK: Elsevier.

Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality*, *9*, 227–245.

Sidman, M. (1960). *Tactics of experimental research: Evaluating experimental data in psychological research*. New York: Basic Books.

Skinner, B. F. (1963). Operant behavior. *American Psychologist*, *18*, 503–515.

SPSS, Inc. (2005). *SPSS for Windows*.

Tawney, J. W., & Gast, D. L. (1984). *Single-subject research in special education*. Columbus, OH: Merrill.

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Lawrence Erlbaum Associates.

Whitehurst, G. (2005). *The Institute of Education Sciences*: *New wine, new bottles*. U.S. Department of Education, Institute of Education Sciences. Retrieved December 13, 2005, at http://www.ed.gov/rschstat/research/pubs/ies.html

Yeaton, W. H. (1982). A critique of the effectiveness of applied behavior analysis research. *Advances in behavior research and therapy*, *4*, 75–96.

This page is left intentionally blank

VALIDATION OF COGNITIVE OPERATIONS AND PROCESSES ACROSS ABILITY LEVELS AND INDIVIDUAL TEST ITEMS

Dimiter M. Dimitrov

ABSTRACT

Knowledge about cognitive operations and processes (COPs) required for success (1 = correct, 0 = incorrect) on test items or learning tasks is very important for in-depth understanding of the nature of student performance and the development of valid instruments for its measurement. A key problem in obtaining such knowledge is the validation of hypothesized COPs and their role in the measurement properties of test items. To provide validation feedback for both normally achieving students and students with learning disabilities, it is important to obtain information on the validity of the COPs for students at different ability levels and individual test items (or tasks). To address this issue, the present chapter introduces a method of estimating the probability for correct performance on individual COPs at fixed ability levels thus providing validity information across ability levels and individual test items. When item response theory (IRT) estimates of the item parameters are known (e.g., in a test bank of IRT calibrated items or published

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 55–81

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19003-5

research), the proposed validation method does not require information about raw (or ability) scores of examinees. This method is illustrated for algebra test items and reading comprehension test items calibrated in IRT.

The cognitive structure of a test is typically defined as a set of cognitive operations and processes (*COP*s), as well as their relationships, required for obtaining correct answers of the test items (e.g., Gitomer & Rock, 1993; Riley & Greeno, 1988). Researchers in learning and cognition have always been challenged with measuring *COP*s that cannot be directly observed. Knowledge about latent *COP*s underlying the student success on test items or learning tasks is important for better understanding the nature of a specific proficiency and the development of valid instruments. Such knowledge can also help educators in developing teaching strategies that target specific cognitive and processing criteria for normally achieving students and/or students with learning disabilities (LD).

Relating test items (or learning tasks) to *COP*s is typically inferred from a theoretical model of knowledge and cognition (e.g., Embretson, 1995; Irvine & Kyllonen, 2002; Mislevy, 1993; Nichols, Chipman, & Brennan, 1995; Snow & Lohman, 1984). For example, Embretson (1995) proposed seven cognitive operations required for correct solutions of mathematical word problems by operationalizing three types of knowledge – *factual* (or *linguistic*), *schematic*, and *strategic* – defined in the theoretical model of Mayer, Larkin, and Kadane (1984). In a learning disability study (Lucangeli, Tressoldi, & De Candia, 2005), a didactic curriculum was developed to give teachers educational strategies useful for improving cognitive processes in six areas: *counting*, *lexical processes*, *semantic processes*, *syntactic processes*, *oral calculation*, and *written calculation*. Target population in this study were students with developmental dyscalculia – a learning disability defined by difficulties on (a) understanding base concepts of arithmetic operations, (b) processing operational symbols, (c) manipulating standard arithmetic concepts, (d) recording data in mathematical problem-solving, (e) calculations, and (f) learning basic arithmetic facts.

The validation of cognitive structures is a key problem and involves productive integration of cognitive psychology and psychometric modeling. Some previous studies have integrated cognitive structures of tests with item response theory (IRT) models for the prediction of item difficulty from cognitive and processing operations (e.g., Embretson, 1984, 1995). Most frequently, this has been done with the validation of cognitive structures for item difficulty prediction using the linear logistic test model (LLTM) in Rasch measurement (e.g., Embretson & Wetzel, 1987; Fischer, 1973; Spada & Kluwe, 1980; Spada & McGaw, 1985; Whitely & Schneider, 1981). An approach to validating cognitive subordinations among test items has been developed in the framework of structural equation modeling (Dimitrov & Raykov, 2003). Other studies have proposed cognitive diagnostic models that bring together cognitive psychology and psychometrics with focus on cognitive error diagnosis, task analysis, and pattern classifications (e.g., DiBello, Stout, & Roussos, 1995; Hartz, 2002; Henson & Douglas, 2005; Junker & Sijtsma, 2001; Maris, 1999; Samejima, 1995; Tatsuoka, 1985, 1995; Tatsuoka, Corter, & Tatsuoka, 2004; Tatsuoka & Tatsuoka, 1987; Tatsuoka & Ferguson, 2003).

Different methods, each with their advantages and disadvantages, provide different perspectives in cognitive validation and analysis. For example, LLTM validation tests (e.g., Fischer, 1973, 1995; Medina-Diaz, 1993) target accuracy in the prediction of item difficulty from *COP*s, but they do not tap into cognitive relationships among items and do not provide cognitive diagnostic information. Recently developed cognitive diagnostic models (e.g., Junker & Sijtsma, 2001; Henson & Douglas, 2005; Tatsuoka, 1985, 1995; Tatsuoka & Ferguson, 2003) provide information about students' profiles on mastering a set of latent *COP*s using more sophisticated (and technically more complex) probabilistic modeling within the framework of both parametric and nonparametric IRT models. Regardless of their aspects and level of theoretical and practical merit, all previous methods of cognitive validation and diagnosis (a) require information about the examinees' scores on individual test items, and (b) do not provide explicit information on cognitive validity across separate items and examinees' ability levels. In an attempt to address this issue, this chapter introduces an approach to validation of *COP*s, required for the correct solution of binary items (or tasks), across individual items and fixed ability levels. The proposed method does not require any information about raw (or ability) scores of examinees, as long as IRT estimates of the item parameters are available. This can be very useful, for example, in (a) validation screening of *COP*s related to IRT bank items, without being necessary to administer such items to examinees, or (b) providing additional perspectives on validation results from previous studies that report IRT estimates of the items parameters. To better understand the theoretical framework of the method introduced in this chapter, as well as its advantages and limitations, the introduction of some basic IRT concepts is necessary.

ITEM RESPONSE THEORY CONCEPTS

This section introduces some basic IRT concepts used in the theoretical framework for validation of *COP*s presented in this chapter. For more information on IRT concepts and models, the reader may refer to IRT textbooks (e.g., Bond & Fox, 2001; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Wright & Stone, 1979).

Under IRT, the term *ability* connotes a latent trait (proficiency, affect) that underlies the responses of examinees on the items of an instrument. The location of a person on the continuum of this latent trait is his/her *ability score*. The units of the ability scale, called *logits*, typically range from −4 to 4. They represent the natural logarithm of the odds for success on the test items. For example, if a person succeeds on 80% and fails on 20% of the test items, the odds ratio for the success on the test is 4/1 = 4. Thus, the ability score of this person is the natural logarithm of 4, which is 1.4 (rounded to the nearest tenth), i.e., about one and a half units above the origin of the logit scale when the person ability distribution is centered at the origin (zero) of the scale.

The graph of the relationship between the ability score of a person and the probability that this person will answer correctly a specific item is called the *item characteristic curve* (*ICC*). Fig. 1 provides the *ICC* for two hypothetical binary items. As one can see, the probability for correct answer on Item 1 is 0.5 (50% chances for correct answer) for persons with ability score of 1.0. Therefore, by the IRT definition of *item difficulty* (denoted *b*), the difficulty of Item 1 is 1.0 on the logit scale ($b_1 = 1$). The difficulty of Item 2, then, is −1.0 ($b_2 = -1$) because the probability for correct answer is 0.5 for persons located at –1.0 on the logit scale. Also, persons with ability score of zero on the logit scale have a probability of 0.38 (38% chances) for success on Item 1 and a probability of 0.82 (82% chances) for success on Item 2. However, persons with ability scores below −2.0 (e.g., some LD students in standardized testing) have better chances for success on Item 1 than on Item 2. Therefore, although the item difficulty parameter of Item 1 ($b_1 = 1$) is greater than this of Item 2 ($b_2 = -1$), Item 1 is only relatively more difficult than Item 2 because the *ICC*s of the two items cross.

While Item 1 is more difficult than Item 2 for persons with ability score of zero, Item 2 better discriminates persons with abilities close to zero, because the *ICC* for Item 2 is steeper than it is for Item 1 at zero. In fact, Item 2 works better than Item 1 in discriminating persons with abilities close to any point in the interval, say, from –3 to 1. Conversely, Item 1 discriminates better than Item 2 in the interval, say, from 1 to 3. In IRT, the steepness of the *ICC* slope for an item at its difficulty location on the scale is measured by

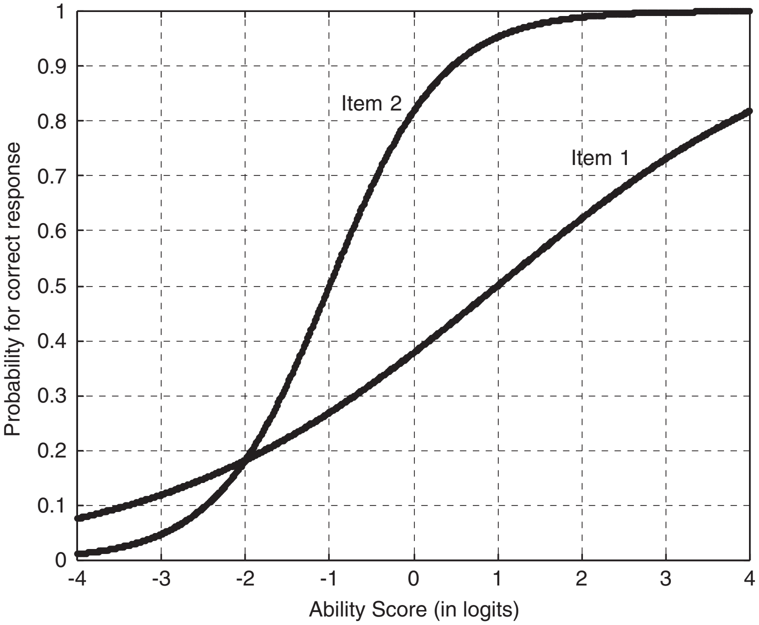

Fig. 1. Item Characteristic Curves (ICCs) for two Items: Item 1 ($a_1 = 0.5$, $b_1 = 1.0$) and Item 2 ($a_2 = 1.5$, $b_2 = -1.0$).

its *discrimination index* (denoted *a*). The item discrimination index typically ranges from 0 to 2. The higher the discrimination index, the steeper the *ICC* at the location of the item difficulty. In Fig. 1, the *ICC* of Item 1 and Item 2 are developed with discrimination indices $a_1 = 0.5$ and $a_2 = 1.5$, respectively.

One-Parameter Logistic Model

If the *ICC*s for all items in a test are almost parallel, then the item discrimination index will be about the same and thus can be fixed. Therefore, the difficulty of an item is the only parameter that governs the probability for success on this item for a person with a given ability score. This is the case of the *one-parameter logistic model* (1PLM) referred to also as the Rasch model (Rasch, 1960/1992). To learn more about the Rasch model, the reader may refer, for example, to Wright and Stone (1979), Bond and Fox (2001), and a series of articles under the section "Understanding Rasch

Measurement" of the *Journal of Applied Measurement*. With the Rasch model for binary items (1 = correct, 0 = incorrect), the probability for correct answer on item i with difficulty b_i for a person with ability score θ (on the logit scale) is

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \tag{1}$$

It is important to note that *person ability* (θ) and *item difficulty* (b) are measured on the same (logit) scale. When $\theta = b$ in Eq. (1), $\exp(\theta - b) = \exp(0) = 1$ and, thus, the probability for correct item response is $P_i(\theta) = 0.5$. In other words, when the ability score for a person and the difficulty parameter for an item share the same location on the logit scale, this person has 50% chances to answer the item correctly. It can be also seen from Eq. (1) that $P_i(\theta) > 0.5$ when $\theta > b$ and, conversely, $P_i(\theta) < 0.5$, when $\theta < b$.

Two-Parameter Logistic Model

When not all items have the same index of discrimination (i.e., not all *ICC*s are parallel; see Fig. 1), then two item parameters (*discrimination* and *difficulty*) will govern the probability for success on any item for a person with a given ability score. This is the case of the *two-parameter logistic model* (2PLM). Under this model, the probability for correct answer on item i, with index of discrimination a_i and difficulty b_i, for a person with ability score θ on the logit scale is

$$P_i(\theta) = \frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]} \tag{2}$$

where D is a constant referred to as *scaling factor*. A typical choice for the scaling factor is 1.7, as it has been shown that when $D = 1.7$, the values for $P_i(\theta)$ for the 2PLM and the two-parameter normal ogive model differ in less than 0.01 in absolute value (Haley, 1952). As with the Rasch model, Eq. (2) produces (a) $P_i(\theta) = 0.5$ when $\theta = b$, (b) $P_i(\theta) > 0.5$ when $\theta > b$, and (c) $P_i(\theta) < 0.5$ when $\theta < b$ (see, Fig. 1).

Three-Parameter Logistic Model

The Rasch model and the two-parameter logistic model assume no guessing in the persons' responses on test items. If guessing is involved, then a third (guessing) item parameter comes into play in addition to item discrimination and item difficulty. In this case the probability for item success is determined with the *three-parameter logistic model* (3PLM). Specifically, the probability for correct response on item *i*, with discrimination a_i, difficulty b_i, and "guessing" index c_i, for a person with ability score θ on the logit scale is

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]} \tag{3}$$

where D is the scaling factor ($D = 1.7$). It has been shown (Lord, 1974) that the values of the "guessing" parameter, c_i, are usually smaller than the values that would result from random guessing – for example, $c_i < 0.20$ for a multiple choice item with five options because the probability for a correct answer on such item with random guessing is 0.20 (1 out of 5). For this reason, the "guessing" parameter, c_i, is referred to in IRT as *pseudo-chance level* parameter.

Eqs. (1)–(3) show that the probability for correct item response depends on the person's ability score, θ, and the item parameters: b_i with the 1PLM (Rasch), a_i and b_i with the 2PLM, and a_i, b_i, and c_i with the 3PLM. Thus, when IRT estimates of the item parameters are available (e.g., in a bank of IRT calibrated items), one can estimate the probability for correct item response for persons at a fixed ability level. This IRT feature is used in the method of validating cognitive operations/processes for students across different ability levels (e.g., LD and normally achieving students on binary test items or learning tasks) presented in the next section.

METHOD

Let COP_1, COP_2, … , COP_m denote the *COP*s related to a test of *n* binary items, and the $n \times m$ matrix of "weights" for these operations is $\mathbf{W} = (w_{ik})$, where $w_{ik} = 1$ if the correct response on item *i* requires correct performance on the cognitive operation COP_k and $w_{ik} = 0$, otherwise. It is assumed that the correct performance on cognitive operations required by an item is statistically independent for a person at a given ability level. With this, the probability of correct response on item *i* for a person *j* with ability θ_j, denoted P_{ij}, relates to the probabilities for this person to perform correctly the cognitive operations required by the item as follows:

$$P_{ij} = \prod_{k=1}^{m} [P_{ij}(COP_k|\theta_j)]^{w_{ik}} \tag{4}$$

where $P_{ij}(COP_k|\theta_\mathrm{j})$ is the person's probability to perform correctly COP_k ($k = 1, \ldots, m$). Taking the natural logarithm on both sides of Eq. (4), we have

$$\ln\ P_{ij} = \sum_{k=1}^{m} w_{ik} \ln\ P_{ij}(COP_k|\theta_j) \tag{5}$$

Given the IRT calibration of binary items, P_{ij} is estimated with the IRT model that has been used in the IRT calibration – for example, Rasch model (Eq. (1)), 2PLM (Eq. (2)), or 3PLM (Eq. (3)). With this, ln P_{ij} in the left side of Eq. (5) is known and, therefore, this equation will produce a system of n linear equations with m unknowns, ln $P_{ij}(COP_k|\theta_j)$, for a fixed ability level, θ_j. The matrix algebra form of this system of linear equations is

$$\mathbf{L} = \mathbf{W} \cdot \mathbf{X} \tag{6}$$

where:

L is the ($n \times 1$) vector with elements ln P_{ij} (known),

W is the ($n \times m$) matrix of weights, w_{ik}, and

X is the ($m \times 1$) vector with unknown elements: ln $P_{ij}(COP_k|\theta_j)$; ($i = 1, \ldots, n$; $k = 1, \ldots, m$).

In general, the system of linear equations in Eq. (6) does not have exact solutions because it is overdetermined – the number of equations is greater than the number of unknowns ($n > m$). Also, the set of *COP*s does not account for item "uniqueness" not captured in matrix **W** and/or aberrant examinees' behavior (e.g., "guessing" or "slipping") in performing *COP*s. Therefore, for a fixed ability level, θ_j, the system of linear equations is solved here using a least squares method of approximation. Specifically, solutions are sought that minimize the *Euclidean norm* of the vector $\|\mathbf{W} \cdot \mathbf{X} - \mathbf{L}\|$ using the *least squares distance* (LSD) method in the context of matrix algebra. As the probability of correct *COP* performance is always between 0 and 1, $0 < P_{ij}(COP_k|\theta_j) < 1$, the elements of the solution vector **X** should be restricted to negative numbers: $\ln P_{ij}(COP_k|\theta_j) < 0$. It should also be reminded that the *Euclidean norm* of a vector is the square root of the sum of the squared elements of the vector. For a comprehensive treatment of solving least squares problems with matrix algebra, the reader may refer, for example, to Lawson and Hanson (1974).

For a fixed ability level, θ_j, the LSD for the norm $||\mathbf{W} \cdot \mathbf{X} - \mathbf{L}||$ is calculated along with the solutions $X_k = \ln P_{ij}(COP_k|\theta_j)$ which represent the elements of vector **X**. Given X_k, the probability for an examinee with ability θ_j to process correctly COP_k is: $P_{ij}(COP_k|\theta_j) = \exp(X_k)$. This makes it possible then to develop a *probability curve* for COP_k across different ability levels. Note that an *ICC* (see Fig. 1) represents the probability for a correct response on the item, whereas the *probability curve* for COP_k is the probability for correct performance on the COP_k across different ability levels (logit scores).

The proposed method is referred here as LSD method (LSDM) for diagnostic validation of *COP*s. A preliminary technical check on results obtained with the LSDM is to examine how close are the LSD values for the solution vector **X** with elements restricted to negative numbers to those for the (theoretically better) solution vector **X** with no sign restrictions on its elements. Relatively large discrepancies between the LSD values under these two scenarios at a given ability level may question the validity of LSDM results at that ability level. With this, kept in mind, the LSDM results translate into three basic *validation criteria* (VC):

- VC_1 (small LSD values): The smaller the LSD value at a given ability level, the better the *COP*s are expected to hold at this ability level.
- VC_2 (monotonicity): There should be a monotonic relationship between ability levels and the probability for correct performance on individual *COP*s – for example, higher reading ability should lead to higher chances for correct performance on a *COP* representing a reading subskill.
- VC_3 (LSDM recovery of *ICC*s): The closer the estimates of the two sides in Eq. (4) for an item at a fixed ability level, the better the *COP*s should hold for the item at this ability level. Graphically, this can be evaluated by the degree of LSDM "recovery" of the *ICC* – that is, the degree of fit between the *ICC* and the line connecting the dots that represent the product of LSDM estimates of probabilities for correct performance on the *COP*s required by the item (e.g., see Figs. 4 and 7).

The validation criteria VC_1, VC_2, and VC_3 should be analyzed collectively as they provide different perspectives in the validation of *COP*s. For example, VC_1 and VC_2 relate to validity of the *COP*s for all items together, whereas VC_3 provides validity information by individual items. In the two illustrative examples that follow, the LSDM is conducted in four steps:

Step 1: The first step is to select an IRT model that fits the data – starting with the simplest, 1PLM (Rasch) model, and then 2PLM and 3PLM, if necessary.

Step 2: The second step is to select (fix) ability levels within a reasonable range on the logit scale. To match students abilities to item difficulties, the range of selected ability levels should cover the interval of item difficulties on the logit scale.

Step 3: For each of the fixed ability levels, the probability for correct item response is estimated with Eqs. (1), (2), or (3), depending on which IRT model (Rasch, 2PLM, or 3PLM) fits the data. By replacing the natural logarithm of this probability, ln P_{ij}, for the term in the left-hand side in Eq. (5), we obtain a system of linear equations, with unknowns ln $P_{ij}(COP_k|\theta_j)$, for a fixed ability level.

Step 4: The system of linear equations generated with Eq. (5) is solved separately for each of the fixed ability levels by minimizing the norm $||\mathbf{W}\cdot\mathbf{X}-\mathbf{L}||$ with the LSD method. The calculations are performed using a computer program in MATLAB (MathWorks, Inc., 1999) developed by the author for practical applications of the LSDM.

Example 1

In this example, the LSDM is illustrated with the validation of a hypothesized cognitive structure for an algebra test of 15 items (simple linear equations) for a sample of 278 ninth-grade high-school students. The appendix provides the algebra test and seven *COP*s required for the correct solutions of the test items. The examinees were required to "show work" on each item, thus avoiding random guessing, and each item was binary scored (1 = correct, 0 = incorrect). The weight matrix **W** (for mapping items to the hypothesized *COP*s) is given in Table 1.

RESULTS

First, it was found that the Rasch model fits the data – binary scores of 278 students on 15 algebra items. Specifically, the conditional likelihood-ratio test (Andersen, 1973), reported with the computer program LPCM-WIN 1.0 (Fischer & Ponochny-Seliger, 1998), indicated a reasonable data fit for the Rasch model, $\chi^2(14) = 22.83$, $p > 0.05$. Therefore, there was no need to test for data fit with the 2PLM or 3PLM. The estimates of the Rasch item difficulty parameter, b, are presented in Table 2. As can be seen, Item 8 is the easiest item ($b_8 = -2.4871$) and Item 11 is the most difficult item ($b_{11} = 2.3913$) in the algebra test of 15 linear equations.

In this example, 13 ability levels were used to cover the interval from -3.0 to 3.0, with an increment of 0.5 ($\theta_j = -3, -2.5, \ldots, 2.5, 3.0$) on the logit scale. The LSD values that minimize the norm $||\mathbf{W}\cdot\mathbf{X}-\mathbf{L}||$ at these ability levels are plotted in Fig. 2, for the cases when the elements of the solution

Table 1. Matrix **W** For the Algebra Test Items and Seven *COP*s.

Item	Cognitive Operation/Process (*COP*)						
	COP_1	COP_2	COP_3	COP_4	COP_5	COP_6	COP_7
I_1	0	0	1	1	1	0	0
I_2	0	0	1	1	0	0	0
I_3	1	1	0	1	1	0	0
I_4	1	0	1	1	1	0	0
I_5	1	0	0	0	0	0	1
I_6	1	1	1	1	0	0	0
I_7	1	0	1	1	1	1	0
I_8	1	0	0	1	0	0	0
I_9	1	0	1	0	0	0	0
I_{10}	1	0	1	1	0	0	0
I_{11}	1	0	1	1	0	0	1
I_{12}	1	0	1	1	1	0	0
I_{13}	1	0	1	1	1	1	0
I_{14}	1	1	1	1	1	0	0
I_{15}	1	0	1	1	1	0	0

Table 2. Estimates of Rasch Item Difficulty, *b*, for 15 Algebra Test Items.

Item	*b*
1	−1.3050
2	−1.6323
3	1.1704
4	−0.2654
5	0.1923
6	0.7265
7	0.7931
8	−2.4871
9	−1.6323
10	−0.1440
11	2.3913
12	−0.6997
13	0.9262
14	2.0282
15	0.3225

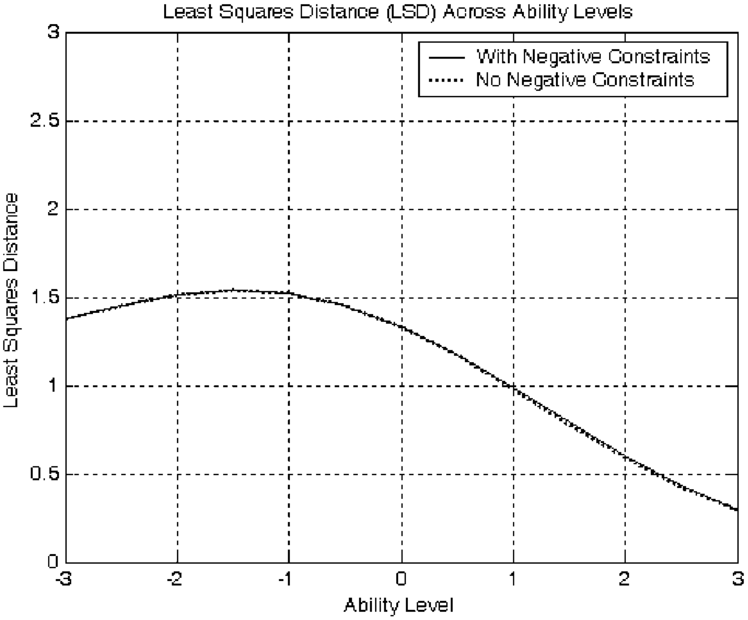

Fig. 2. Least Squares Distance (LSD) Obtained with Minimizing the Norm||**W** · **X**–**L**||across Seven Ability Levels for the Algebra Test: LSDs Obtained with the Solutions in Vector **X** Restricted to Negative Numbers and LSDs with Unrestricted Solutions in Vector **X**.

vector **X** are (a) restricted to negative numbers and (b) not restricted to negative numbers. The LSD values obtained under these two minimization scenarios are almost identical at all ability levels, thus satisfying the initial technical condition for validity of the LSD results across ability levels. One can also notice that smaller LSD values are obtained at high-ability levels. This indicates that the hypothesized *COP*s hold better for high-ability examinees (VC_1).

The LSDM estimates of probabilities for correct performance on each *COP* are reported in Table 3 and plotted in Fig. 3 across the 13 ability levels on the logit scale. The positive monotonic relationship between these probabilities and ability levels is consistent with the second validation criterion, VC_2. The probability curves for the seven *COP*s provide also information about their relative difficulty and discrimination at different ability levels. For example, most difficult across all ability levels are COP_7 ("removing denominators") and COP_2 ("solving for nonnumerical coefficients"),

Table 3. Probability for Correct Processing of Nine Cognitive Operations/Processes for the Algebra Test Items at 13 Ability Levels.

Ability level (θj)	Cognitive Operation/Process (*COP*)						
	COP_1	COP_2	COP_3	COP_4	COP_5	COP_6	COP_7
−3.0	0.5182	0.1441	0.3743	0.4960	0.5718	0.3751	0.0719
−2.5	0.6112	0.1539	0.4319	0.5623	0.6059	0.3734	0.0862
−2.0	0.7092	0.1691	0.4959	0.6238	0.6482	0.3810	0.1052
−1.5	0.8085	0.1918	0.5631	0.6788	0.6974	0.4003	0.1306
−1.0	0.9033	0.2244	0.6291	0.7274	0.7506	0.4342	0.1640
−0.5	0.9863	0.2699	0.6899	0.7708	0.8034	0.4846	0.2075
0.0	1.000	0.3372	0.7531	0.8256	0.8563	0.5583	0.2718
0.5	1.000	0.4223	0.8083	0.8747	0.9014	0.6446	0.3515
1.0	1.000	0.5201	0.8536	0.9133	0.9363	0.7326	0.4427
1.5	1.000	0.6224	0.8905	0.9417	0.9615	0.8117	0.5412
2.0	1.000	0.7192	0.9205	0.9616	0.9782	0.8748	0.6403
2.5	1.000	0.8019	0.9442	0.9750	0.9885	0.9201	0.7323
3.0	1.000	0.8663	0.9623	0.9839	0.9942	0.9504	0.8106

Note: The entries are $P(COP_k|\theta_j) = \exp(X_k)$, where X_k are the LSD estimates for the solution elements of vector $\mathbf{X}$ in minimizing the norm $\|\mathbf{W} \cdot \mathbf{X} - \mathbf{L}\|$.

whereas the easiest is COP_1 ("solving for a variable with a numeric coefficient"). Also, the examinations of the steepness of the probability curves shows that more difficult *COP*s (COP_7, COP_2, and COP_6) discriminate well among high-ability examinees. Conversely, relatively easy *COP*s (COP_1, COP_2, COP_4, and COP_3) discriminate better among low-ability examinees. This makes perfect sense, given that the performance of high achievers usually varies more on difficult tasks, whereas the performance of low achievers varies more on easy tasks.

Additional diagnostic validation is provided with the degree to which the *COP*s hold for individual items across ability levels (see VC_3). Ideally, the probability for correct response on a given item (in this case, estimated with Eq. (1)) should equal the product of the probabilities for correct processing of the *COP*s required by the item (see Eq. (4)). For example, the **W** matrix in Table 1 shows that the first item ($i = 1$) requires COP_3, COP_4, and COP_5. Taking into account Eq. (4), this translates into the equation $P_{ij} = P(COP_3|\theta_j)P(COP_4|\theta_j)P(COP_5|\theta_j)$ for a fixed ability level ($\theta_j = -3.0$, –2.5, … , 2.5, 3.0) on the logit scale. Table 3 provides estimates of the probabilities for correct performance on each *COP* across items and fixed ability levels. For each item, the absolute difference between the estimates for the two sides in Eq. (4) at each ability level, as well as the *mean absolute*

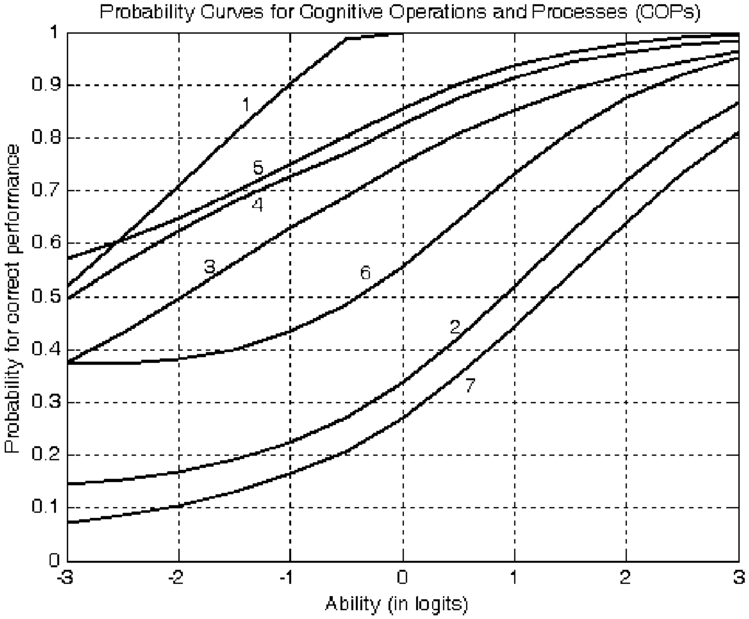

Fig. 3. Probability Curves for Seven COPs Required for the Correct Solution of the Algebra Test Items.

difference (*MAD*) across ability levels, are tabulated in Table 4. Graphically, the LSDM recovery of *ICC*s is illustrated in Fig. 4 for four items (3, 5, 6, and 10). Similar graphs were developed for the other 11 items, but they are not provided here for space consideration. Ideally, $MAD = 0$ would indicate perfect LSDM recovery of the *ICC* for an item. The examination of the 15 graphs and their *MAD* values in Table 4 revealed an *excellent* overall LSDM recovery of the *ICC*s for three items (3, 7, and 13), *good* recovery for one item (4), *somewhat good* recovery for eight items (6, 8, 9, 10, 11, 12, 14, and 15), and *relatively poor* recovery for two items (1 and 5).

Example 2

In this example, the LSDM is illustrated for five *COP*s that were hypothesized to govern the correct response on 10 multiple-choice items in a reading comprehension test. The sample study consisted of 234 high-school juniors. Two essays were used, with each essay composed of three parts: an

Table 4. Absolute Differences Between the Probability for Correct Item Response Estimated with the Rasch Model and the Product of LSDM Estimates of the Probabilities for Correct Processing of the COPs Required by the Items Across Ability Levels for the Algebra Test.

Item	Ability (logits)							
	−3	−2	−1	0	1	2	3	*MAD*
1	0.049	0.132	0.232	0.254	0.179	0.099	0.045	0.150
2	0.017	0.100	0.195	0.215	0.153	0.089	0.044	0.124
3	0.006	0.008	0.008	0.002	0.013	0.020	0.014	0.010
4	0.006	0.008	0.014	0.034	0.050	0.040	0.022	0.026
5	0.020	0.066	0.160	0.276	0.324	0.259	0.150	0.188
6	0.010	0.024	0.058	0.116	0.162	0.145	0.086	0.090
7	0.001	0.004	0.008	0.014	0.017	0.012	0.006	0.009
8	0.117	0.177	0.159	0.098	0.057	0.027	0.012	0.095
9	0.009	0.057	0.085	0.083	0.079	0.054	0.028	0.060
10	0.042	0.084	0.115	0.086	0.021	0.010	0.012	0.055
11	0.002	0.011	0.035	0.085	0.146	0.163	0.120	0.082
12	0.036	0.072	0.115	0.136	0.116	0.071	0.035	0.087
13	0.001	0.003	0.008	0.014	0.016	0.012	0.006	0.009
14	0.001	0.007	0.023	0.063	0.116	0.130	0.090	0.063
15	0.020	0.053	0.100	0.112	0.067	0.023	0.006	0.058

Note: Reported are the absolute differences across seven ability levels, from −3.0 to 3.0 with a "step" of 1.0, but the *mean absolute difference* (*MAD*) is calculated for all 13 ability levels, with a "step" of 0.5 on the logit scale.

introduction, a middle passage consisting of two contrasting positions, and a thesis statement. The essays were designed to be as similar as possible in terms of vocabulary, syntactical complexity, and length – the first essay, "Health Hazards of Smoking," was 208 words, and the second essay, "Heaven's Gate Mass suicide," was 209 words in length. Originally, nine items were written for each essay, but for the illustration with this example, only 10 (out of 18) items were used after eliminating items that were difficult (or inappropriate) to specify with the **W** matrix. This is because the essays were not originally developed for validation of cognitive operations and processes (see Henning, 1999). For the illustration purposes of this example, the following five *COP*s were hypothesized to relate to the choice of correct answer for the 10 essay items:

- COP_1 = Using a word-matching strategy in selecting the correct option,
- COP_2 = Using a number-matching strategy in selecting the correct option,

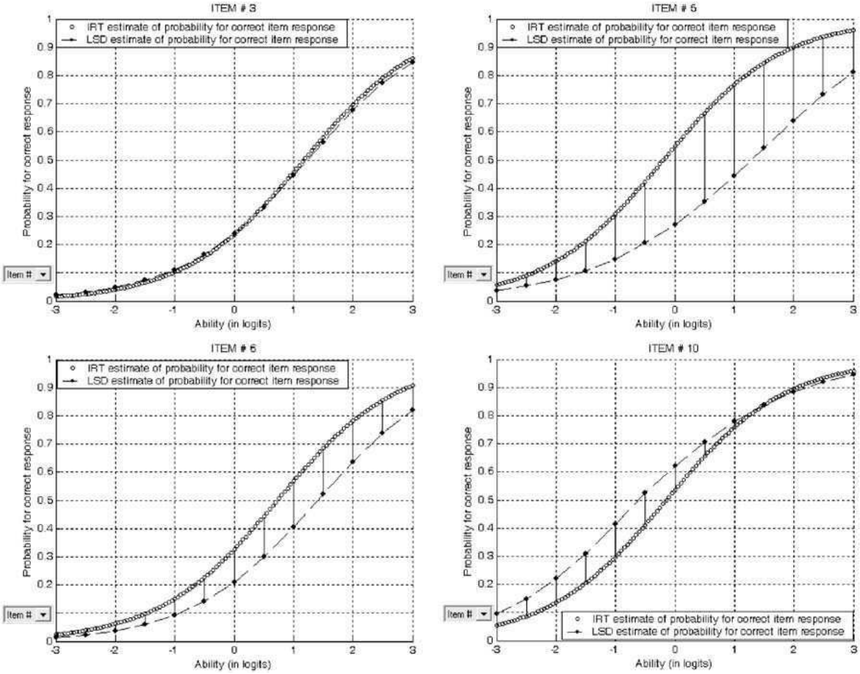

Fig. 4. LSDM Recovery of Item Characteristic Curves (ICCs) for Four Items in the Algebra Test.

- COP_3 = Processing relevant information located in an introductory passage,
- COP_4 = Processing relevant information located in an ending passage, and
- COP_5 = Processing *distractors* (statements that match the wording of the false answer).

These five *COP*s can be viewed as general proxies for some (among dozens) cognitive attributes identified in previous research on difficulties in reading comprehension (e.g., Buck, Tatsuoka, & Kostin, 1997; Embretson & Wetzel, 1987; Perkins, Gupta, & Tammana, 1995). For example, Buck et al. (1997) used 24 cognitive attributes in a study on subskills of reading for the Test of English for International Communication, TOEIC), taken by more than 30,000 test-takers in Japan and Korea. However, given the lower complexity and small number of essay items used in this example, the five *COP*s listed

Table 5. **W** Matrix for Five *COP*s Required by Ten Reading Comprehension Items.

Item	Cognitive Operation/Process				
	COP_1	COP_2	COP_3	COP_4	COP_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	1	0	1	0
4	0	1	0	0	0
5	0	1	1	0	0
6	0	0	0	1	0
7	0	1	0	0	1
8	0	0	0	1	0
9	0	0	0	1	1
10	1	0	1	0	0

here above were considered appropriate for the illustration of diagnostic validation with the LSDM.

The **W** matrix in Table 5 shows the association between the five *COP*s and individual items. The LSDM results were obtained by conducting the four steps described earlier (in the method section) in this chapter. As with Example 1, the computer program in MATLAB (MathWorks, Inc., 1999), developed for calculations and graphics with the LSDM, was used to facilitate the diagnostic validation of *COP*s across ability levels and individual test items in this example.

Results

First, it was checked whether the one-parameter (Rasch) model fits the data – binary scores (1 = true, 0 = false) of 234 students on 10 items of reading comprehension. Unlike Example 1, the Rasch model did not fit the data in this example, as indicated by the conditional likelihood-ratio test reported with the computer program LPCM-WIN 1.0: $\chi^2(11) = 24.64$, $p<0.05$. Therefore, the two-parameter IRT model (2PLM) was tested next for data fit using the computer program XCALIBRE (Assessment System Corporation, 1995). For the test data fit, XCALIBRE reports a standardized residual statistic for each item. This statistics follows (approximately) the standard normal distribution, $N(0,1)$, and values in excess of 2.0 indicate misfit with a Type I error rate of 0.05. In this case, the standardized residuals for the 10 essay items ranged from 0.09 to 1.02 thus indicating that the data

Table 6. IRT Estimates of Item Discrimination (a) and Item Difficulty (b) for Ten Reading Comprehension Items.

Item	a	b
1	0.60	−2.03
2	0.81	−1.29
3	0.75	−1.03
4	0.81	−1.58
5	0.62	0.59
6	0.75	−1.65
7	0.54	2.22
8	0.65	−1.46
9	0.75	2.58
10	0.54	−0.66

fit the 2PLM. The XCALIBRE estimates of the items parameters with the 2PLM, *item discrimination* (a) and *item difficulty* (b), are given in Table 6.

The validation of the hypothesized *COP*s was conducted across 13 fixed ability levels in the interval from −3.0 to 3.0, with an increment of 0.5: $\theta_j = -3, -2.5, \ldots, 2.5, 3.0$ (logits). For each ability level, the probability for correct item response was estimated with the 2PLM (see Eq. (2)) using the item parameter estimates in Table 6; ($D = 1.7$ for the scaling factor). The natural logarithm of this probability, $\ln P_{ij}$, was then used in the left-hand side in Eq. (5) for each of the 10 items, thus generating a system of 10 linear equations with six unknown elements, $\ln P_{ij}(COP_k|\theta_j)$; ($i = 1, \ldots, 10$; $j = 1, \ldots, 13$; $k = 1, \ldots, 5$).

The resulting system of linear equations was solved separately for each of ability level by minimizing the norm $\|\mathbf{W} \cdot \mathbf{X} - \mathbf{L}\|$ with the LSDM. The LSD values are plotted in Fig. 5 for the cases when the elements of vector **X** are (a) restricted to negative numbers and (b) not restricted to negative numbers. As can be seen, the LSD values obtained under these two minimization scenarios are almost identical at all ability levels thus satisfying the initial condition for validity of the LSD results across ability levels. Also, smaller LSD values are obtained at high-ability levels thus indicating that the hypothesized *COP*s should hold better for high-ability examinees.

The LSDM estimates of the probabilities for correct performance on individual *COP*s are reported in Table 7 and plotted in Fig. 6 across ability levels. The monotonic increase of these probabilities across ability levels is yet another piece of evidence in validating the *COP*s in this example. Clearly, COP_5 (processing distractors) is the most difficult cognitive attribute, followed (in decreasing difficulty) by COP_3, COP_4, COP_2, and (the

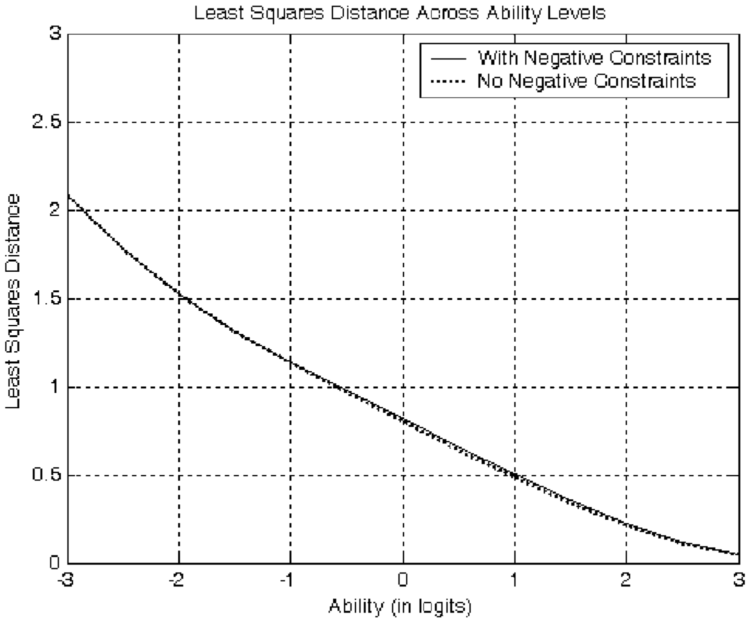

Fig. 5. Least Squares Distance (LSD) Obtained with Minimizing the Norm$||\mathbf{W}\cdot\mathbf{X}-\mathbf{L}||$across Seven Ability Levels for the Reading Comprehension Items: LSDs Obtained with the Solutions in Vector **X** Restricted to Negative Numbers and LSDs with Unrestricted Solutions in Vector **X**.

easiest) COP_1 (using a word-matching strategy in selecting the correct option). Fig. 6 shows, for example, that the probability of correct performance on COP_5 for average ability examinees located at the origin of the logit scale ($\theta = 0$) is about 0.08 (more accurately, 0.0758 as indicated in Table 7). That is, the chances for these examinees to correctly "process distractors" (COP_5) are about eight percent. Also, the steepness of the probability curves is consistent with the known measurement logic that difficult *COP*s discriminate better among high-ability examinees, whereas easy *COP*s discriminate better among low-ability examinees.

For each item, the absolute difference between the estimates for the two sides in Eq. (4) at each ability level, as well as the *MAD* across the ability levels, is provided in Table 8. Graphically, the LSDM recovery of *ICC*s is illustrated in Fig. 7 for four items (1, 3, 7, and 8). Similar graphs were developed for the remaining six items, but they are not provided here for space consideration. The examination of all graphs and associated *MAD*

Table 7. Probability for Correct Processing of Five Cognitive Operations and Processes for the Reading Comprehension Test at 13 Ability Levels.

Ability (θj)	Cognitive Operation/Process (*COP*)				
	COP_1	COP_2	COP_3	COP_4	COP_5
−3.0	0.4178	0.2112	0.1617	0.1461	0.0146
−2.5	0.5828	0.3216	0.1751	0.2187	0.0167
−2.0	0.7635	0.4667	0.1967	0.3131	0.0200
−1.5	0.9303	0.6329	0.2308	0.4235	0.0254
−1.0	1.000	0.7873	0.2906	0.5376	0.0346
−0.5	1.000	0.9052	0.3776	0.6419	0.0501
0.0	1.000	0.9832	0.4794	0.7284	0.0758
0.5	1.000	1.000	0.5952	0.8001	0.1184
1.0	1.000	1.000	0.7058	0.8574	0.1835
1.5	1.000	1.000	0.7974	0.9032	0.2754
2.0	1.000	1.000	0.8664	0.9398	0.3930
2.5	1.000	1.000	0.9146	0.9677	0.5259
3.0	1.000	1.000	0.9466	0.9866	0.6555

Note: The entries are $P(COP_k|\theta_j) = \exp(X_k)$, where X_k are the LSD estimates for the elements of vector $\mathbf{X}$ in minimizing the norm $\|\mathbf{W} \cdot \mathbf{X} - \mathbf{L}\|$; ($k = 1, \ldots, 5$; $j = 1, \ldots, 13$).

values revealed an overall *good* recovery of the *ICC*s for two items (7 and 9), *somewhat good* recovery for three items (3, 5, and 8), *somewhat poor* recovery for two items (4 and 6), and *poor recovery* for two items (1 and 2). This classification, however, is somewhat loose because the recovery differences were much more pronounced at the below average ability levels for several items (e.g., see Fig. 7, items 1 and 3). In general, such diagnostic information can be particularly useful in validating reading skills for students with learning disabilities.

Overall, the results indicate that the five *COP*s used in this example relate to difficulties in reading comprehension but, as expected, they do not represent sufficiently complete and valid set of reading comprehension subskills. As noted at the beginning of this example, they can serve as proxies to some variables identified in more refined models of reading comprehension difficulty (e.g., Embretson & Wetzel, 1987; Buck et al., 1997). With this understanding, and within the illustration purposes of this example, the five *COP*s exhibit logical measurement behavior in terms of monotonicity, relative difficulty, and discrimination (see Fig. 6). This, in combination with the diagnostic validation of *COP*s across individual items, can provide

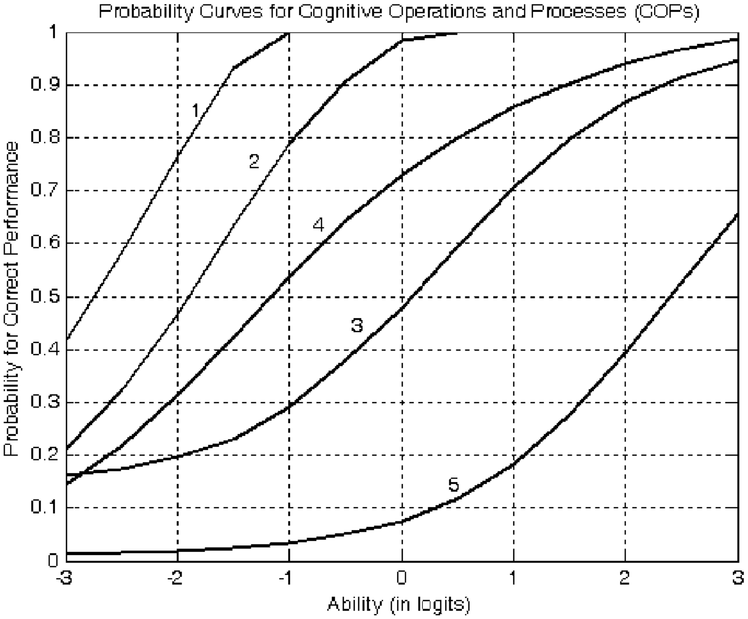

Fig. 6. Probability Curves for Five COPs Required for the Correct Answer of Ten Reading Comprehension Items.

Note: COP_1 = Using a word-matching strategy in selecting the correct option, COP_2 = Using a number-matching strategy in selecting the correct option, COP_3 = Processing relevant information located in an introductory passage, COP_4 = Processing relevant information located in an ending passage, and COP_5 = Processing distractors.

useful feedback for developing more refined sets of reading skills targeted, say, with local assessments of reading comprehension.

SUMMARY

The validation of cognitive structures involves productive integration of cognitive psychology and psychometric modeling. Previous studies have integrated cognitive structures of tests with (a) item response theory (IRT) models to predict item difficulty from cognitive operations and processes (e.g., Embretson, 1995; Fischer, 1995; Whitely, 1980), (b) structural equation modeling to validate cognitive subordinations among test items

Table 8. Absolute Differences Between the Probability for Correct Item Response Estimated with the 2PLM and the Product of LSDM Estimates of the Probabilities for Correct Processing of the *COP*s Required by the Items Across Ability Levels for the Reading Comprehension Test.

Item	Ability (logits)							
	−3	−2	−1	0	1	2	3	*MAD*
1	0.147	0.256	0.259	0.112	0.043	0.016	0.006	0.105
2	0.125	0.194	0.189	0.128	0.041	0.011	0.003	0.078
3	0.044	0.079	0.086	0.072	0.073	0.040	0.008	0.025
4	0.088	0.108	0.097	0.085	0.028	0.007	0.002	0.043
5	0.012	0.031	0.071	0.122	0.099	0.051	0.019	0.039
6	0.005	0.077	0.159	0.163	0.109	0.051	0.012	0.060
7	0.005	0.011	0.022	0.040	0.062	0.057	0.016	0.021
8	0.008	0.042	0.087	0.106	0.081	0.039	0.006	0.036
9	0.001	0.003	0.008	0.019	0.040	0.046	0.016	0.017
10	0.037	0.076	0.132	0.168	0.116	0.054	0.020	0.050

Note: Reported are the absolute differences across seven ability levels, from −3.0 to 3.0 with a "step" of 1.0, but the *mean absolute difference* (*MAD*) is calculated for all 13 ability levels, with a "step" of 0.5 on the logit scale.

(Dimitrov & Raykov, 2003), and (c) parametric and nonparametric IRT models for cognitive error diagnosis, task analysis, and pattern classifications (e.g., DiBello et al., 1995; Henson & Douglas, 2005; Junker & Sijtsma, 2001; Samejima, 1995; Tatsuoka, 1985, 1995; Tatsuoka & Ferguson, 2003).

The method introduced in this chapter provides diagnostic information about the validity of hypothesized *COP*s across different ability levels and individual test items. This method is based on LSD in minimizing the matrix norm $\|\mathbf{W}\cdot\mathbf{X}-\mathbf{L}\|$, where **W** is the weight matrix for mapping individual items to hypothesized *COP*s, **L** the vector of natural logarithms of the probability for correct item response, and **X** the vector of natural logarithms of the probabilities for correct performance on the *COP*s at a fixed ability level, θ_j– that is, the elements of vector **X** are: $X_k = \ln P_{ij}(COP_k|\theta_j)$. With this, the probability for correct performance on a cognitive operation, COP_k, at a fixed ability level, θ_j, is estimated as $P_{ij}(COP_k, |\theta_j) = \exp(X_k)$.

The estimates of the probabilities $P_{ij}(COP_k|\theta_j)$, resulting from LSD solutions in minimizing the matrix norm $\|\mathbf{W}\cdot\mathbf{X}-\mathbf{L}\|$, are then used to develop (a) probability curves for the *COP*s (e.g., Figs. 3 and 6) and (b) LSDM recovery of item characteristic curves (e.g., Figs. 4 and 7). These results are interpreted in light of the validation criteria described earlier in this chapter:

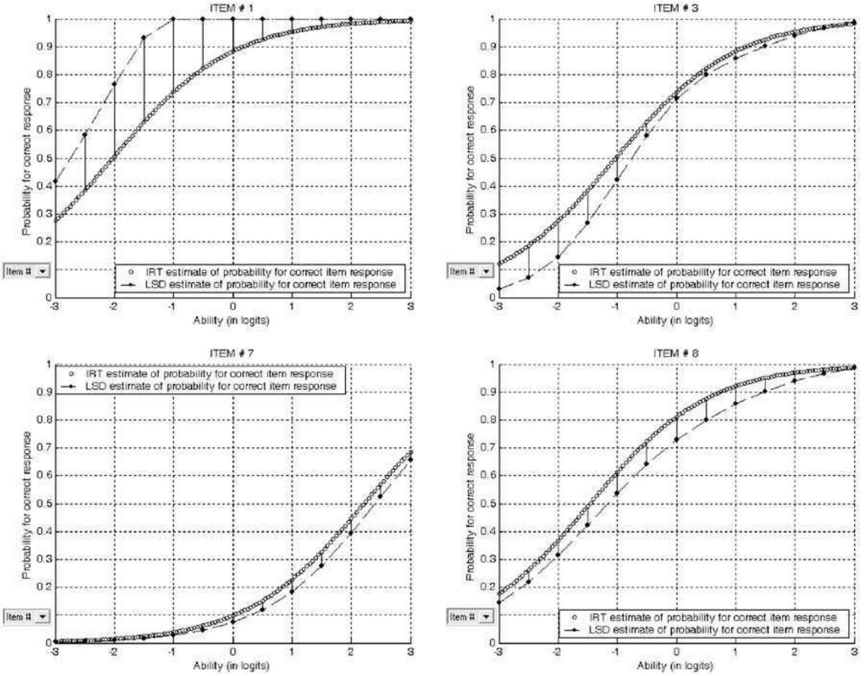

Fig. 7. IRT Item Characteristic Curves and their LSD Approximation for Four Items of the Reading Comprehension Test.

relatively small LSDs; monotonicity of probability curves for *COP*s; and LSDM recovery of *ICC*s. When the LSDM recovery of the *ICC* for some items is not satisfactory, the search for plausible explanations may also contribute to better understanding of relationships between such items and the *COP*s that they require. Technically, the LSDM allows for a quick exploratory search for the presence of *COPs* not initially hypothesized in **W,** but such analysis should be substantively based (e.g., within a specific model of knowledge and cognition).

One limitation of the LSDM is that it does not produce a statistical test (e.g., χ^2 test for goodness-of-fit) for the overall validity of cognitive structures. Another limitation is that the LSDM does not account for factors such as "guessing" or "slipping" that may affect the examinees performance on individual *COP*s. Such factors are incorporated, at the expense of more complicated modeling and technicality, into some recent approaches to cognitively based assessment, test development, error diagnosis, and pattern classifications (e.g., DiBello et al., 1995; Henson & Douglas, 2005; Junker & Sijtsma, 2001; Samejima, 1995; Tatsuoka, 1985; Tatsuoka & Ferguson, 2003). Instead, the LSDM provides diagnostic validation of hypothesized *COP*s across different ability levels and individual test items through deterministic solutions for probabilistic relationships between individual items and their *COP*s (Eq. (4)).

Advantages of the LSD method to other methods of cognitive diagnosis and validation are that the LSDM (a) does not require information about examinees' raw (or ability) scores, as long as IRT estimates of the item parameters are available, and (b) provides diagnostic information on *COP*s validity across fixed ability levels and individual items. One practical implication is that, using the LSDM, researchers would be able to validate *COP*s related to IRT bank items, without administering such items to examinees. In another situation, researchers would be able to "cross-validate" and provide additional perspectives on cognitive validation results from, say, previous studies that report IRT estimates of the items parameters.

In conclusion, the LSDM approach to diagnostic validation of *COP*s reveals important aspects of their measurement behavior across ability levels and individual items (or tasks) and may have valuable applications in the research on cognition and learning in diverse settings.

REFERENCES

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Assessment System Corporation (1995). *User's manual for XCALIBRE marginal maximum-likelihood estimation program (Windows version 1.0)*. St. Paul, MN: Author.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Buck, G., Tatsuoka, K. K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*(3), 423–466.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In: P. Nichol, S. Chipman & R. Brennan (Eds), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Erlbaum.

Dimitrov, D. M., & Raykov, T. (2003). Validation of cognitive structures: A structural equation modeling approach. *Multivariate Behavioral Research*, *38*(1), 1–23.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Embretson, S. E. (1995). A measurement model for linking individual learning to process and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, *32*, 277–294.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*, 175–193.

Fischer, G. (1995). The linear logistic model. In: G. H. Fischer & I. W. Molenaar (Eds), *Rasch models: Foundations, recent developments, and applications* (pp. 131–155). NY: Springer-Verlag.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G., & Ponochny-Seliger, E. (1998). *Structural Rasch modeling: Handbook of the usage of LPCM-WIN 1.0*. Groningen, The Netherlands: ProGAMMA.

Gitomer, D. H., & Rock, D. (1993). Addressing process variables in test analysis. In: N. Frederiksen, J. Mislevy & I. Bejar (Eds), *Test theory for a new generation of tests* (pp. 243–268). Hillsdale, NJ: Erlbaum.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error*. Stanford, CA: Stanford University, Applied Mathematics and Statistics Laboratory Technical Report No 15.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation. University of Illinois, Urbana-Champain.

Henning, J. E. (1999). *Signs of development in written composition: Identifying thinking in the expository essays of eleventh graders*. Unpublished doctoral dissertation. Kent State University, Kent, Ohio.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*, 262–277.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Lawson, C. L., & Hanson, R. J. (1974). *Solving least squares problems*. Englewood Cliffs, NJ: Prentice-Hall.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264.

Lucangeli, D., Tressoldi, P., & De Candia, C. (2005). Education and treatment of calculation abilities of low-achieving students and students with dyscalculia: Whole class and individual implementations. In: T. E. Scruggs & M. A. Mastropieri (Eds), *Advances in learning and behavioral disabilities: Cognition and learning in diverse settings* (Vol. 18, pp. 199–223). Oxford, UK: Elsevier.

Maris, E. (1999). Estimating multiple classification latent class models. *Pscyhometrika*, *64*, 187–212.

MathWorks, Inc. (1999). *Learning MATLAB (version 6.0)*. Natick, MA: Author.

Mayer, R., Larkin, J., & Kadane, P. (1984). A cognitive analysis of mathematical problem solving. In: R. Stenberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 231–273). Hillsdale, NJ: Erlbaum.

Medina-Diaz, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement*, *17*, 117–130.

Mislevy, R. J. (1993). Foundations of a new theory. In: N. Frederiksen, R. J. Mislevy & I. Bejar (Eds), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale, NJ: Erlbaum.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial network. In: A. Davies & J. Upshur (Eds), Language testing. (Vol.12, Issue 1, pp. 34–53). London: Edward Arnold.

Rasch, G. (19601992). *Probabilistic models for some intelligence tests*. Chicago: MESA Press (Original work published 1960).

Riley, M. S., & Greeno, J. C. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction*, *5*, 49–101.

Samejima, F. (1995). A cognitive diagnosis method using latent trait models: Competency space approach and its relationship with DiBello and Stout's unified cognitive-psychometric diagnosis model. In: P. Nichol, S. Chipman & R. Brennan (Eds), *Cognitively diagnostic assessment* (pp. 391–410). Hillsdale, NJ: Erlbaum.

Snow, R. E., & Lohman, D. F. (1984). Implications of cognitive psychology for educational measurement. In: R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.

Spada, H., & Kluwe, R. (1980). Two models of intellectual development and their reference to the theory of Piage. In: R. Kluwe & H. Spada (Eds), *Developmental model of thinking* (pp. 1–32). New York: Academic Press.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In: S. Embretson (Ed.), *Test design: New directions in psychology and psychometrics* (pp. 169–193). New York: Academic Press.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, *10*, 55–73.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In: P. Nichol, S. Chipman & R. Brennan (Eds), *Cognitively diagnostic assessment* (pp. 327–360). Hillsdale, NJ: Erlbaum.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of Royal Statistical Society*, *65*, 143–157.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, *4*, 901–926.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and statistical pattern classification. *Psychometrika*, *52*(2), 193–206.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*(4), 479–494.

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, *5*, 383–397.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

APPENDIX. TEST OF ALGEBRA LINEAR EQUATIONS

1. $2(6x+3) = 3x-10$
2. $5x-3 = 7+10x$
3. $2x-n = 5$
4. $5(2x-5x)-12 = 20x$
5. $3x/7 = 2$
6. $nx-a = 5a$
7. $4[x+3(x-2)] = 10$
8. $-7-2x = 10$
9. $20 = 5-5x$
10. $2x+7-10x+3 = 12-2x$
11. $4x/5+2+2x/3-10 = 0$
12. $-5(8-2x) = 2x-2$
13. $5-2(x+3) = x+5(2x-1)+10$
14. $-4(5x-4)+5x = 10-n+2n$
15. $-6(x-4)+2x = 5x-10$

Cognitive operations/processes (*COP*s) required for the correct solution of the algebra test equations:

COP_1: Solving for variable with a numeric coefficient
Example: If $5x = 20$ then $x = 20/5$

COP_2: Operating with non-numerical coefficients
Example: If $nx = a+5$ then $x = (a+5)/n$

COP_3: Collecting terms
Example: If $2x-5x+8x = 10$ then $5x = 10$

COP_4: Balancing
Example: If $5+2x = 3$ then $2x = 3-5$

COP_5: Removing parentheses
Example: If $5(2x+3) = 5$ then $10x+15 = 5$

COP_6: Removing brackets (or more parentheses)
Example: If $4[x+3(x-2)] = 10$ then $16x-24 = 10$

COP_7: Removing denominators
Example: If $3x/7 = 1/2$ then $6x = 7$

This page is left intentionally blank

STRUCTURAL EQUATION MODELING: “RULES OF THUMB” WITH PARTICIPANTS WITH DISABILITIES

Giulia Balboni

ABSTRACT

Structural Equating Modeling (SEM) is a formal model for representing dependency relations between variables of psychological events and may be used for verifying the structural organization of a theoretical model. “Rules of thumb” for the use of SEM are presented regarding each step of its application: specification of the structural model, measurement of the psychological event, and estimation of the adequacy of the model in representing the event. The investigation of the factorial structure of Greenspan’s model of personal competence is presented as an example of SEM application with participants with disabilities.

Structural Equating Modeling (SEM) is a formal model for representing dependency relations between variables of psychological events (e.g., Jöreskog, 1973; McDonald, 1999). SEM may be used to identify a structural model of interrelated variables which may explain, i.e., arguably causal, an observed

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 83–100

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19004-7

and measured psychological event. More specifically, SEM may be used to verify the structural organization of a theoretical model, to reveal how the different factors of a theoretical model are organized. In this case, we have a model of Confirmative Factor Analysis (CFA), a special type of SEM.

SEM is based on the hypothesis that, given an empirical variance–covariance matrix between the observed variable of a psychological event, it is possible to identify an adequacy structural model. That is, a model whose parameters produce an estimated population variance–covariance matrix which is close to the sample matrix (e.g., Bentler & Weeks, 1980; Jöreskog & Sörbom, 1996). An estimated matrix is close to a sample matrix when there are no statistically significant differences among them, when the discrepancies among them are minimal, according to indexes of goodness-of-fit (e.g., χ^2, Root Mean Square Error of Approximation, RMSEA; McDonald, 1999).

The term SEM, also known as path analysis with latent variables, is also used to describe the data analysis method for representing dependency relations in multivariate data (e.g., Bagozzi, 1979; Bollen, 1989; Maruyama, 1998).

The SEM data analysis method requires different steps (e.g., McDonald & Ho, 2002):

1. Specification of the structural model, that is, to hypothesize the variables and the relationships between them that may explain the psychological event. The variables may be constructs as well as observed variables, e.g., intelligence and scores on an intelligence test. The relation may be of dependency, e.g., the construct of intelligence influences scores in the intelligence test, as well as of covariance, e.g., the relationship between intelligence and another construct, e.g., adaptive behavior. Relations between the variables are representing with a system of linear regression equations.
2. Measurement of the psychological event that the structural model should explain. For this purpose, a variance–covariance matrix should be developed between the observed variables, indicators of the constructs hypothesized in the structural model.
3. Empirical estimation of the adequacy of the structural model in describing the measured psychological event. A series of goodness-of-fit indices must be calculated. The purpose is to investigate the closeness of the estimated variance–covariance matrix (based on the structural model) to the observed variance–covariance matrix. Eventually, modification of the structural model to increment its goodness-of-fit.

4. Comparison of the structural model (original or modified version) with alternative versions of the model.
5. Evaluation of the estimated parameters of the structural model final version, e.g., the regression coefficients of the dependency relations.

The purpose of SEM is to describe a psychological event via the simplest structural model (e.g., Jöreskog, 1973; Markus, 1998; Raykov & Marcoulides, 1999). SEM may be used to identify, between the different structural models hypothesized to describe a psychological event, the one that is adequate and at the same time the most parsimonious. That means that, comparing alternative structural models with the same goodness-of-fit, the most parsimonious model must be chosen.

For any set of multivariate data, measurement of the psychological event, there will almost always be more than one plausible structural model (e.g., Jöreskog, 1973). Given a structural model that is parsimonious and adequate in describing a psychological event, researchers may infer only that the model has not been falsified but not that it has been verified. Other structural models could describe the data with the same or greater level of appropriateness.

The purpose of this chapter is to describe the steps required to apply SEM. In particular, the application of SEM given a model of CFA will be explained. In addition, an empirical example of CFA with participants with mental retardation will be presented: an investigation of the factorial structure of Greenspan's model of personal competence (Greenspan & Driscoll, 1997).

PHASE 1. SPECIFICATION OF THE STRUCTURAL MODEL

Given a psychological event to explain, first, a structural model of interrelated variables must be hypothesized.

The structural model is made up of variables and parameters. Variables may be latent or observed. Latent variables are unobserved variables, theoretical constructs. Observed variables are manifest variables, indicators of the constructs. Both latent and observed variables may be exogenous and endogenous. Exogenous variables are independent, that is, variables whose values do not depend on any other variables of the model. Endogenous variables are dependent, that is, variables whose values depend on at least one other variable of the model.

Parameters of the model are: (1) structural coefficients that represent the dependency relations between independent and dependent variables; (2)

covariance coefficients that represent the inter-relation between independent variables; and (3) variance of independent variables. The errors are represented via independent variables. The errors are measurement errors, if they are observed variables, and specification error, if they are latent variables. Parameters may be free, when their value must be estimated, or fixed, when their value is fixed by the researcher at a given value.

The part of the model that relates the observed variables to the corresponding latent variables is generally called the measurement model. The hypothesized relationships among the latent variables are called the path model. The term structural model refers to the combined measurement and path model (McDonald & Ho, 2002).

A structural model is graphically represented via a path diagram (e.g., Pearl, 2000) and formally represented via a system of linear regression equations.

Confirmative Factor Analysis

Figure 1 represents an example of a path diagram in the case of CFA. There is only a measurement model and not a path model – variables are only independent. Variables are latent, i.e., factors of the theoretical model (ξ), and observed, i.e., indicators of each factor (x), and indicators' measurement errors (δ). Parameters to be estimated are structural coefficients, i.e., factor loadings of each indicator with respect to the corresponding factor (λ); covariance coefficients, i.e., covariance between factors (ϕ) and between

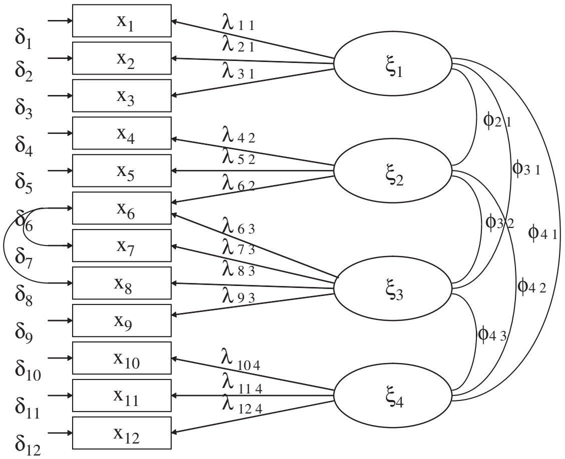

Fig. 1. Path Diagram of a CFA Structural Model.

measurement errors (indicated with an arc among the measurement errors); and variance of independent variables measurement errors.

To scale factors, factor variance is fixed as equal to one. To estimate measurement errors, the corresponding structural coefficients is fixed as equal to one (e.g., Bentler, 1995; Byrne, 1994; Jöreskog & Sörbom, 1996).

The system of regression equations which describe the measurement model are made up of one equation for each indicator. Example of an equation is given for the first indicator:

$$X_1 = \lambda_{11}\xi_{11} + \delta_1$$

Conditions for identification of the structural model. The structural model must be identified, that is, there must be a unique numerical solution for each parameter of the model. For this purpose, the number of parameters to be estimated must be less than the data points, i.e., the number of non-redundant variances and covariances in the sample matrix. Moreover, to identify a CFA model with interrelated factors, each factor must has at least two pure indicators, i.e., indicator that loads on only one factor (McDonald, 1999).

PHASE 2. MEASUREMENT OF THE PSYCHOLOGICAL EVENT

Given a structural model, observed variables hypothesized, i.e., indicators in the case of CFA, must be measured. An empirical variance–covariance matrix between the observed variables must be produced.

To use SEM, the measured variables must satisfy statistical assumptions. Observed variables must be measured with at least an interval scale and be multivariate normally distributed. Investigations about the robustness of the multivariate normality assumption have found that parameter estimates remain valid under reasonable assumptions even when the data are non-normal, whereas standard errors do not (e.g., Satorra & Bentler, 1994; Hu & Bentler, 1995). SEM methods for analysis of observed covariance matrices are available for use with non-interval or non-normal variables, e.g., asymptotically distribution-free estimators (Browne, 1984), or continuous/categorical variable methodology estimators (Muthén, 1984). Both estimators required very large sample size, larger than those generally expected.

SEM is a large-sample technique. Generally, a sample size of about 200 is adequate for small to medium models (Boomsma, 1983). Regarding the relation among numbers of participants and of parameters, it seems that there should be almost five participants for each parameter to be estimated, if the observed variables are normally distributed, or almost 10 participants, if the normally assumption is not satisfied (Bentler & Chou, 1987). Recently, it was suggested that, with normally distributed observed variables, fewer than 10 participants per estimated parameter may be adequate if the model explained a big portion of the observed variables variance (MacCallum, Brown, & Sugawara, 1996; Ullman, 2001).

Finally, dependency as well as covariance relations among variables must be linear and participants must be independent to each others.

PHASE 3. ESTIMATION OF ADEQUACY OF THE STRUCTURAL MODEL AND ITS MODIFICATION

Estimation of Structural Model Adequacy

Given a structural model of a psychological event and a measurement of a psychological event, the adequacy of the structural model in representing the psychological event must be estimated. For this purpose, different statistical packages are available, e.g., LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), and AMOS (Arbuckle, 1997). Estimation methods must be used, e.g., Maximum Likelihood, Asymptotically Distribution Free.

First, sample data are used to estimate parameters of the model. Then, estimated parameters are used to produce the estimated population variance–covariance matrix to be compared to the observed sample matrix. The degree of closeness between estimated and observed matrixes allow for the evaluation of the adequacy of the structural model. For this purpose, it is necessary to: (a) check the identification of the model, (b) evaluate the indices of goodness-of-fit, and (c) analyze the residuals between estimated and observed covariance matrixes.

(a) Identification of the model. A model must be identified, i.e., the solution must be unique. For this purpose, as it has been already explained, conditions must be satisfied in the specification of the structural model. However, even when these conditions are satisfied, the model may not be identified. Examples of indications of non-identification are excessively large standard errors of the parameters, and Heywood solutions, e.g., variance of measurement errors negative or close to zero, or standardized structural or covariance coefficients greater than 1 (e.g., Ullman, 2001).

(b) Indices of goodness-of-fit. If the model seems to be identified, then it is necessary to evaluate indices of goodness-of-fit, i.e., discrepancies among

estimated and observed variance–covariance matrixes. A good fit is sometimes indicated by a non-significant χ^2. However, χ^2 may be not a good index. For example, with large sample, trivial differences between observed and estimated matrices are often significant, even if the model is adequate (Bentler, 1988).[1] Therefore, several indices have been developed (e.g., Tanaka, 1993). They may be distinguished relative and absolute indices.

Relative indices compare a function of the discrepancy from the fitted model to a function of the discrepancies from then null model (generally, a model that corresponds to completely unrelated variables). An example is the Tucker Lewis Index (TLI; Tuker & Lewis, 1973). The TLI has a range of zero to one, with value equal or greater to 0.90 indicating a good fit.

Absolute indices are functions of the discrepancy (and sometimes of the sample sizes and number of parameters to be estimated). Examples of more frequently used absolute indices are the Adjusted Goodness-of-Fit Index (AGFI), the RMSEA, and the Standardized Root Mean square Residual (SRMR). AGFI is a measure of the weighted portion of variance in the observed covariance matrix accounted for by the estimated covariance matrix[2] (Tanaka & Huba, 1989). It may assume values between zero and one, and a value of almost 0.90 is suggested. The RMSEA is a measure of approximation of the estimated matrix to the observed matrix (Browne & Cudeck, 1993). Generally, RMSEA less than 0.05 corresponds to a "good" fit and less than 0.08 corresponds to an "acceptable" fit. The SRMR is a measure of the standardized residual between the estimated and observed variance–covariance matrixes. Small values indicate good-fitting models; values of 0.08 or less are desired.

There are several problems with goodness-of-fit indices (McDonald & Ho, 2002). There is no established empirical or mathematical basis for their use. Moreover, there is no sufficiently strong correspondence between alternative indices for a decision based on one to be consistent with a decision based on another. Therefore, some authors have suggested to evaluate adequacy on the model on the bases of two types of indices, e.g., SRMR and another index (Hu & Bentler, 1999).

(c) Examine residuals. In general, a given degree of global misfit can originate from a correctable misspecification giving a few large discrepancies or it can be due to a general scatter of discrepancies not associated to any particular misspecification. Thus, it is necessary to examine the standardized residual covariance matrix, the discrepancies between estimated and observed covariance matrixes. In this way, correctable misspecification can be identified, i.e., large residuals not symmetric, and therefore variable relations may be added.

In the case of CFA, standardized residuals different from zero may represent parameters that have not been hypothesized in the structural model. For example, an indicator hypothesized as loading on only one factor may result in loading on two factors; correlation between measurement errors not specified may result necessary.

Modification of the Specified Model

On the basis of analysis of discrepancies between estimated and observed covariance matrixes, the structural model may be modified to improve its fit. More specifically, statistical tests are available to reveal which modifications in the model specification may improve its fit: Lagrange Multiplier (LM) and Wald tests. The LM test may be used to investigate which parameters should be added to the model to improve the fit. The Wald test may be used to reveal which parameters could be dilated (e.g., Ullman, 2001).

However, a model modification must be theoretically founded (McDonald & Ho, 2002). It is incorrect to add or dilate parameters only on the bases of residuals, LM or Wald tests, if modifications are not theoretically plausible.

Given a modified structural model, it is necessary to compare its goodness-of-fit to the original version of the model. For this purpose, if models are nested, i.e., models are subtests of each other, $\Delta\chi^2$ may be evaluated. The χ^2 value for the larger model is subtracted from the χ^2 value for the smaller nested model and the difference, $\Delta\chi^2$, also a χ^2, is evaluated with degrees of freedom equal to the difference between the degrees of freedom in the two models. A non-statistically significant $\Delta\chi^2$ means that the two models have the same goodness-of-fit; therefore, the smaller nested model, being more parsimonious, must be chosen. A significant $\Delta\chi^2$ means that the nested model, although is more parsimonious, makes worse the model; therefore, the lager model must be chosen.

PHASE 4. COMPARISON OF THE STRUCTURAL MODEL WITH ALTERNATIVE MODEL VERSIONS

The estimation of the adequacy of the structural model allows researchers to infer if the model has not been falsified but not if it has been verified. For any set of multivariate data, there will almost always be more than one plausible structural model. Therefore, given a structural model with a good fit, it is necessary to specify alternative versions of the model, which are theoretically plausible, to compare to the good fit model. The comparison may be based on the evaluation of the goodness-of-fit indices and on $\Delta\chi^2$.

In the case of the CFA, alternative models may have factors, which are the collapsing of pairs of factors of the good fit model.

PHASE 5. EVALUATION OF ESTIMATED PARAMETERS OF THE FINAL VERSION STRUCTURAL MODEL

Given a non-falsified structural model, better than other theoretically plausible versions of the model, the standardized estimated parameters of the model may be evaluated. In the case of CFA, the factor loading coefficients of indicators should be statistically significant and sufficiently large to justify the specification of the indicators loading on the factors. The factor correlation coefficients should not be high enough to justify the collapsing of any pair of factors.

EMPIRICAL EXAMPLE: FACTORIAL STRUCTURE OF GREENSPAN'S MODEL OF PERSONAL COMPETENCE

Factorial structure of Greenspan's model of personal competence was investigated via CFA (Balboni, 2003). Greenspan has proposed a personal competence model that incorporates maladaptive behavior and three distinct types of intelligence: conceptual, social, and practical.

The validity of Greenspan's model has received some empirical support (e.g., McGrew & Bruininks, 1990; McGrew, Bruininks, & Johnson, 1996). However, the participants of the investigations were generally heterogeneous regarding the type of disability, e.g., learning disabilities, mental retardation, and mental illness (McGrew et al., 1996). Moreover, it has been hypothesized that Greenspan's personal competence construct may not be robust across participants ability levels. Thus, the validity of Greenspan's model was investigated in participants homogenous regard the level of ability/disability, i.e., moderate mental retardation.

For this purpose, whether the hypothesized factorial structure of the Greenspan's model was not falsified in children with moderate mental retardation was investigated. A structural model was specified which represented the factor structure of Greenspan's model. Indicators of each factor were measured and goodness-of-fit of the hypothesized factorial model were estimated.

Phase 1. Specification of CFA Structural Model

Fig. 2 represents the path diagram of the hypothesized structural model. As can be seen, it is an empirical example of the theoretical CFA structural model represented in Fig. 1. Factors were the dimensions hypothesized by Greenspan: conceptual, practical, social intelligences, and maladaptive behavior. Indicators of conceptual intelligence were the three Kaufman (1979) factors of *WISC-R* (Wechsler, 1974; 1987 Italian adaptation): Verbal Comprehension, Perceptive Organization, and Concentration. Indicators of

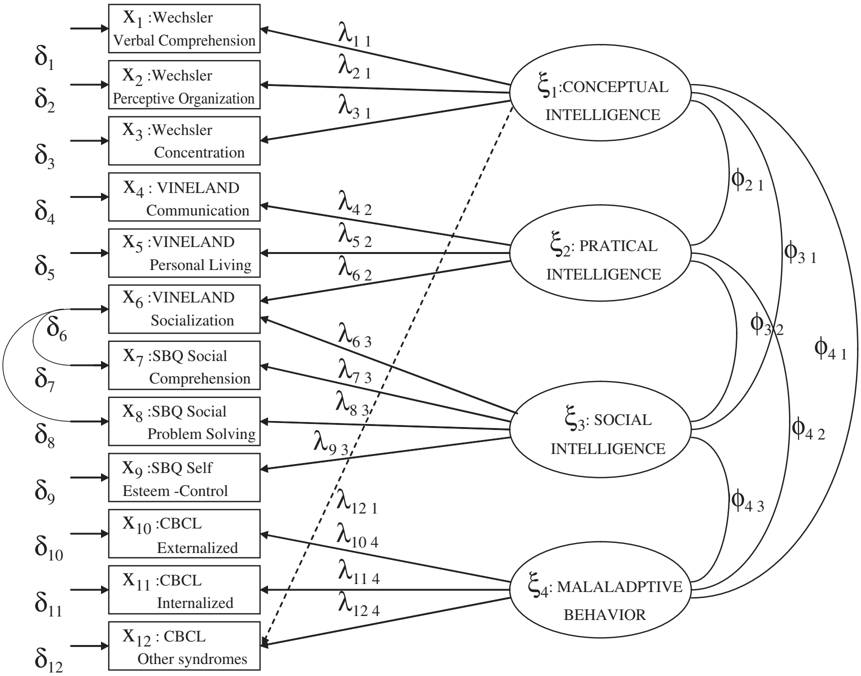
δ1
δ2
δ3
δ4
δ5
δ6
δ7
δ8
δ9
δ10
δ11
δ12
X1 :Wechsler Verbal Comprehension
X2 :Wechsler Perceptive Organization
X3 :Wechsler Concentration
X4 : VINELAND Communication
X5 :VINELAND Personal Living
X6 :VINELAND Socialization
X7 :SBQ Social Comprehension
X8 :SBQ Social Problem Solving
X9 :SBQ Self Esteem -Control
X10 :CBCL Externalized
X11 :CBCL Internalized
X12 : CBCL Other syndromes
λ1 1
λ2 1
λ3 1
λ4 2
λ5 2
λ6 2
λ6 3
λ7 3
λ8 3
λ9 3
λ12 1
λ10 4
λ11 4
λ12 4
ξ1:CONCEPTUAL INTELLIGENCE
ξ2: PRATICAL INTELLIGENCE
ξ3: SOCIAL INTELLIGENCE
ξ4: MALALADPTIVE BEHAVIOR
φ2 1
φ3 1
φ3 2
φ4 1
φ4 2
φ4 3

Fig. 2. Path Diagram of the Structural Model of the Original and Modified Versions of Greenspan's Model of Personal Competence. In Broken Line is the Modification of the Original Model.

practical intelligence were three *Vineland Adaptive Behavior Scales-Survey Form* (Sparrow, Balla, & Cicchetti, 1984; 2003 Italian adaptation): Communication, Daily Living Skills, and Socialization. Indicators of social intelligence were the Vineland Socialization Scale and three items clusters of a Social Behavior Questionnaire (SBC; Adapted by Kevin McGrew (1996) from Selected Scales of Independent Behavior and The Checklist of Adaptive Living Skills; Morreau & Bruininks, 1989; McGrew et al., 1996): Social Comprehension, Social Problem Solving, Self-esteem and Control. Indicators of maladaptive behavior were the three maladaptive behavior scales of the *Child Behavior Check List/4–18* (CBCL; Achenbach, 1991; 1998 Italian adaptation): Externalized, Internalized, and Other Syndromes.

Covariance relations were hypothesized among each possible pair of factors. Moreover, covariance relations between the measurements errors of Vineland Socialization scale and SBQ Social Comprehension and SBQ Social Problem Solving items clusters were hypothesized because items measure similar sample behaviors.

Conditions for identification of the hypothesized structural model were satisfied. To scale the factors, the factors variance were fixed equal to 1. To estimate the measurement errors, the corresponding structural coefficients were fixed equal to 1 (e.g., Bentler, 1995; Byrne, 1994; Jöreskog & Sörbom, 1996). In this way, as can be seen in Fig. 2, the parameters to be estimated were equal to 33, less than the data points, equal to 78.[3] Moreover, each factors had almost two pure indicators.

Phase 2. Measurement of the Factor Indicators

Participants of the investigation were 112 Italian children (65% male) with a diagnosis of moderate mental retardation. They were 6–12-years old (Mean [SD] = 9–0 [1–7]). All the participants lived with their family (95%) or in institutions for children with disabilities (5%); they were included in regular classes (95%) or attended special schools (5%).

The participants had been selected among all the clients of several facilities located in northern and central Italy. One hundred and seventeen participants were initially selected. Selection criteria included (a) caregivers' permission; (b) age between 6 and 12 years; (c) a diagnosis of moderate mental retardation (in agreement with DSM-IV); and (d) had attained a non-zero standard score in at least three verbal and three non-verbal WISC-R sub-tests (1993 Italian standardization; Orsini, 1993). Five participants were eliminated as they were found to be outliers in almost one of the scales used as Greenspan's model factors indicators.

Each participant was tested on the 12 indicators of Greenspan's model with WISC-R, Vineland Scales, SBC, and CBCL (see Fig. 2). To avoid order effects, the order compilation of the four instruments was counterbalanced. To avoid participant and administrator expectations compromising the validity of the investigation, participants as well as administrators were not informed of the research hypothesis.

The SEM assumptions were satisfied. All scores attained in the indicators were interval scale, normally distributed variables (skewness: range $= -0.38$ to 0.60, Median $= 0.17$; kurtosis: range $= -0.81$ to 0.95, Median $= -0.39$). Regarding the sample size, it was less than the proposed value of 200 (Boomsma, 1983). Moreover, there were at least three, but not five, participants for parameters to be estimated (critical value suggested for normally distributed variables; Bentler & Chou, 1987). On the other hand, in agreement with MacCallum, Brown, and Sugawara (1996), the model may be valid if it explains a large portion of the variance of the observed variables. Finally, criteria followed for the selection of participants must satisfy the assumption of independence.

Phase 3. Evaluation of the Adequacy of the Structural Model and its Modification

Via a Maximum Likelihood estimation technique, the specified model was estimated and its adequacy was investigated. As can be seen in Table 1 (original version), the goodness-of-fit was quite good. However, consider that a model with standardized factor loading equal to one could not be identified.

Standardized residual matrix examination revealed some large residuals, i.e., greater than 1.96; in particular, there was a large residual covariance ($z = 2.04$) among Wechsler Concentration (X_3) and CBCL Other Syndromes (X_{12}) indicators. CBCL Other Syndromes scale regards disorder like thought and attention problems, which in some way, are related with sample behavior measured by Wechsler Scale. Therefore, a model modification was specified with CBCL Other Syndromes scale loading conceptual intelligence (Fig. 2, broken line).

Via Maximum Likelihood estimation technique, the modified structural model was estimated. There were no more indices of non-identification. Goodness-of-fit were good (Table 1, modified version). χ^2 were not significant, RMSEA were less than 0.05 ($p = 0.53$),[4] SRMR were less than 0.08, and TLI were more than 0.90. Only the AGFI were not completely satisfying,

Table 1. Statistical Goodness-of-Fit Indexes of the Different Versions of Estimated Greenspan's Model.

	$\Delta\chi^2$	RMSEA	SRMR	TLI	AGFI	χ^2	
						Value	d.f.
Original version	–	0.054	0.07	0.96	0.86	59.44	45
Modified version	4.80*	0.047	0.07	0.97	0.87	54.64	44
Alternative modified version							
1. Collapsing conceptual, practical, and social intelligences	98.28**	0.140	0.10	0.81	0.71	152.92**	50
2. Collapsing conceptual and practical intelligences	56.16**	0.110	0.09	0.87	0.76	110.80**	47
3. Collapsing conceptual and social intelligences	67.61**	0.120	0.10	0.84	0.74	122.25**	47
4. Collapsing practical and social intelligences	33.70**	0.087	0.08	0.91	0.81	88.34**	48

Note: $\Delta\chi^2$ for the modified version were calculated with regard to the original version while for the three alternative versions were calculated with regard to the modified version.

$^{*}p \leqslant 0.05$.

$^{**}p \leqslant 0.001$.

i.e., almost equal to 0.90. Moreover, $\Delta\chi^2$ obtained in the comparison between modified and original model versions were statistically significant; it indicates that the larger model, i.e., the modified model version, was better than the smaller model, i.e., the original model version. Finally the median of standardized residual was equal to zero. Thus, it can be said that the modified version of Greenspan's personal competence model was non-falsified.

Phase 4. Comparison with Alternative Versions of the Structural Model

The estimation of the adequacy of the structural model allows it to be inferred if the model has not been falsified, but not if it has been verified.

Therefore, theoretical plausible alternative versions of the modified model, i.e., good fit model, were specified.

Specifically, four alternative versions of the modified model were specified which included the collapsed factors of the modified model. The first alternative model included the three different intelligences collapsed into a single intelligence factor. The other three alternative models had collapsed all possible pairs of intelligence factors: conceptual and practical intelligences, second model; conceptual and social intelligences, third model; practical and social intelligences, fourth model. These alternative models were specified because investigations (McGrew et al., 1996) revealed that correlations between conceptual, practical, and social intelligences in participants with moderate-severe disabilities are moderate to high, and higher than those of participants with mild disabilities. Therefore, in participants with moderate mental retardation, models with a single collapsing intelligence factor, or two collapsing intelligence factors, could be plausible and alternative to the original three intelligence factors proposed by Greenspan.

Via Maximum Likelihood estimation technique, the goodness-of-fit of the alternative modified model versions were estimated. As can be seen in Table 1, the modified model was better than all its alternative versions. The goodness-of-fit indices of the modified model were better than those of the alternative versions in all cases. Moreover, $\Delta\chi^2$ values were always statistically significant, indicating that the larger model, i.e., the modified model, was better than the smaller model, i.e., each alternative modified model version.

However, the alternative model collapsing practical and social intelligences, even if it was worse than the modified model, had some goodness-of-fit indexes that are quite satisfying, and that suggest it may be a good model. This is an example of the plausible event that a non-falsified model, in this case, the alternative model collapsing practical and social intelligences, may not necessarily be the best model. Given a good fit model, an alternative version must be specified to compare to it.

Phase 5. Evaluation of Estimated Parameters of the Final Version Structural Model

The modified version of Greenspan's model of personal competence has not been falsified and has produced results better than theoretically plausible alternative model versions. Therefore, the estimated parameters of the model are reported in Fig. 3. The factor loading coefficients of all the 12 indicators were statistically significant and the magnitudes were generally

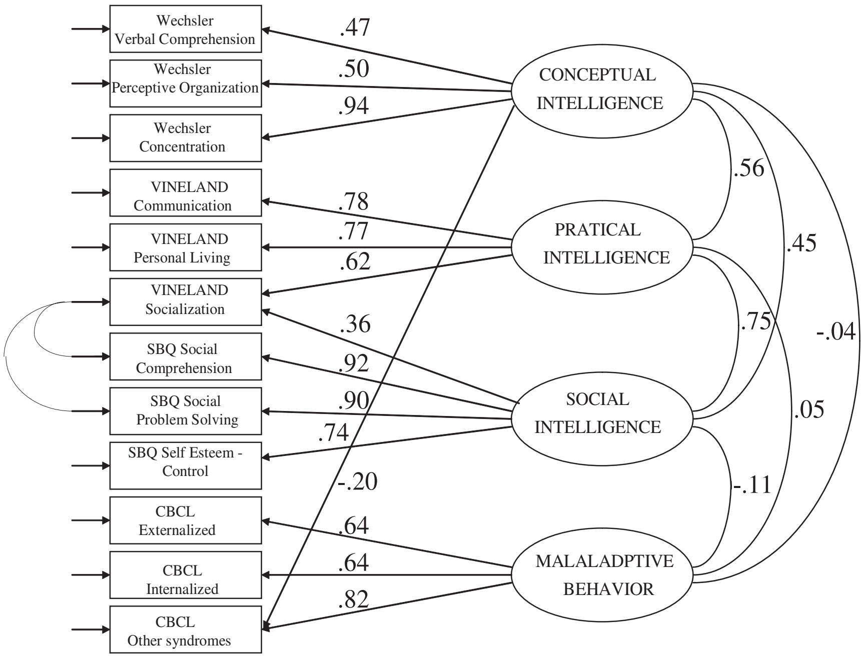

Fig. 3. Standardized Parameters of Greenspan's Model of Personal Competence Estimated in Participant with Moderate Mental Retardation.

moderate to high. Thus, the specification of the indicators loading on the factors was justified. The factor correlation coefficients between the three intelligence factors were statistically significant, and with moderate to high magnitude. The discriminant validity of the model was verified: values of the confidence interval of all correlation coefficients were less than one, value that should indicate the collapsing of factors in a single dimension.

The correlation between the maladaptive behavior and the intelligence factors was not statistically significant. It seems that, in participants with moderate mental retardation, maladaptive behavior factor is not related to intelligence. However, there is a statistically significant factor loading of a maladaptive behavior indicator, Other syndromes, on the conceptual intelligence factor; this may represent an indirect relation between conceptual intelligence and maladaptive behavior.

To reveal the variance explained by the structural model, the squared multiple correlations for each indicators were calculated; they were quite high (range: 0.22–0.88; Median = 0.60) indicating that the model explained quite a large portion of the measured variables. As the model explains such a large portion of the variance of the observed variables, in accordance with MacCallum et al. (1996), the model may be valid, even though the sample size is less than the suggested size.

Factorial structure of Greenspan's model of personal competence was empirically verified via CFA in children with moderate mental retardation. The hypothesized model was not falsified and yielded better results than other theoretically plausible model versions. The constructs of conceptual, social, and practical intelligence as well as affective competence can be said to be distinct dimensions.

Future investigations are needed to cross-validate the results, with a larger sample of participants with moderate mental retardation, and to generalize the results, with participants with different levels of ability/disability and age.

CONCLUSION

CFA, an application of SEM, may be used for verifying the structural organization of a theoretical model. "Rules of thumb" for the use of this method of data analysis have been presented regarding the steps required: specification of the structural model; measurement of the psychological event; estimation of the adequacy of the model in representing the event and, if necessary, its modification; comparison with theoretically plausible alternative version of the model; and evaluation of the estimated parameters of the final version model. An empirical example of CFA with participants with disabilities was presented: the investigation of the factorial structure of Greenspan's model of personal competence with participants with moderate mental retardation. It was shown that, although in the case of participants with disabilities it is difficult to have a large sample size, SEM may be very useful for representing dependency relations between variables of psychological events.

NOTES

1. However, a frequently used "rule of thumb" to verify if the model is adequate is that the ratio of the χ^2 value and its degree of freedom must be less than 2.
2. AGFI is adjusted for the number of parameters estimated in the model.
3. Number of data points (p^*) may be calculated with: $p^* = [p(p+1)]/2$; p = number of observed variables.
4. p-value for test of close fit RMSEA < 0.05.

ACKNOWLEDGMENT

The investigation of Greenspan's model of personal competence was supported in part by a grant (Progetto di Ricerca d'Ateneo 98, Università degli Studi di Padova) awarded to Luigi Pedrabissi by University of Padua, Italy. The author thanks Luigi Pedrabissi for the helpful advice in the realization of the research.

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry. [A. Frigerio (Ed., 1998). *Adattamento Italiano*. Bosisio Parini (LC): Istituto Scientifico E. Medea, Ass. La Nostra Famiglia]

Arbuckle, J. L. (1997). *AMOS users' guide version 3.6*. Chicago: Small Waters.

Bagozzi, R. P. (1979). *Causal models in marketing*. New York: Wiley.

Balboni, G. (2003, May). *Greenspan's model of personal competence: Validity in children with mild and moderate mental retardation*. Paper presented at the annual meeting of the American Academy on Mental Retardation, Chicago.

Bentler, P. M. (1988). Comparative fit indexes in structural model. *Psychological Bulletin*, *107*, 238–246.

Bentler, P. M. (1995). *EQS Structural Equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods and Research*, *16*, 78–117.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equation with latent variables. *Psychometrika*, *45*, 289–308.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Browne, M. W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In: K. A. Bollen & J. S. Long (Eds), *Testing structural equation models* (pp. 136–162). Thousand Oaks, CA: Sage.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Thousand Oaks, CA: Sage.

Greenspan, S., & Driscoll, J. (1997). The role of intelligence in a broad model of personal competence. In: D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 131–150). New York: Guilford Press.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In: R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Hu, L.-T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In: A. S. Goldberger & O. D. Duncan (Eds), *Structural equation models in the social sciences* (pp. 85–112). New York: Seminar Press.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: Users' reference guide*. Chicago: Scientific Software International.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York: Wiley.

MacCallum, R. C., Brown, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Bulletin*, *119*, 130–149.

Markus, K. A. (1998). Judging rules. *The Journal of Experimental Education*, *66*, 261–265.

Maruyama, G. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.

McGrew, K. S., & Bruininks, R. H. (1990). Defining adaptive and maladaptive behavior within a model of personal competence. *School Psychology Review*, *19*, 53–73.

McGrew, K. S., Bruininks, R. H., & Johnson, D. R. (1996). Confirmatory factor analytic investigation of Greenspan's model of personal competence. *American Journal of Mental Retardation*, *100*, 533–545.

Morreau, L. E., & Bruininks, R. H. (1989). *Checklist of adaptive living skills*. Chicago: Riverside.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Orsini, A. (1993). *WISC-R Contributo alla taratura italiana*. Firenze, Italy: Organizzazioni Speciali.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Raykov, T., & Marcoulides, G. A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling*, *6*, 292–300.

Satorra, A., & Bentler, P. M. (1994). Corrections to standard errors in covariance structure analysis. In: A. von Eye & C. C. Clogg (Eds), *Latent variable analysis: Application to developmental research* (pp. 339–419). Thousand Oaks, CA: Sage.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland adaptive behavior scales*. Circle Pines, MN: American Guidance Service [Balboni, G., & Pedrabissi, L. (2003). *Adattamento italiano*. Firenze, Italy: Organizzazioni Speciali].

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In: K. A. Bollen & J. S. Long (Eds), *Testing structural equation models*. Newbury Park, CA: Sage.

Tanaka, J. S., & Huba, G. J. (1989). A general coefficient of determination for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology*, *42*, 233–239.

Tuker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychomtrika*, *38*, 1–10.

Ullman, J. B. (2001). Structural equation modeling. In: B. G. Tabachnick & L. S. Fidell (Eds), *Using multivariate statistics* (4th ed., pp. 653–771). New York: Harper-Collins.

Wechsler, D. (1974). *Wechsler intelligence scale for children-revised*. New York: The Psychological Corporation [Rubini, V., & Padovani, F. (1987). Adattamento italiano, Scala d'Intelligenza Wechsler per Bambini Revisionata. Firenze, Italy: Organizzazioni Speciali].

MODERN ALTERNATIVES FOR DEALING WITH MISSING DATA IN SPECIAL EDUCATION RESEARCH

Craig Enders, Samantha Dietz, Marjorie Montague and Jennifer Dixon

ABSTRACT

Missing data are a pervasive problem in special education research. The purpose of this chapter is to provide researchers with an overview of two "modern" alternatives for handling missing data, full information maximum likelihood (FIML) and multiple imputation (MI). These techniques are currently considered to be the methodological "state of the art", and generally provide more accurate parameter estimates than the traditional methods that are still common in published educational studies. The chapter begins with an overview of missing data theory, and provides brief descriptions of some traditional missing data techniques and their requisite assumptions. Detailed descriptions of FIML and MI are given, and the chapter concludes with an analytic example from a longitudinal study of depression.

Missing data are a common problem in educational research. Factors such as student mobility and socioeconomic status (SES) are constant threats when considering attrition in educational studies, but working with special

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 101–129

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19005-9

education populations poses even more challenges. For example, it has been suggested that absence rates and suspensions for special education students are significantly higher than those of the general education population (e.g., Shriner & Wehby, 2004), making it difficult to obtain complete data from these children. In a similar vein, the dropout rate for special education students may be nearly twice as high as that of general education students (Blackorby & Wagner, 1996; Wagner, Blackorby, Cameto, & Newman, 1993). Clearly, unique characteristics such as these make missing data a near certainty when conducting research with at risk populations.

Unfortunately, researchers have traditionally relied on ad hoc methods for dealing with missing data (e.g., listwise and pairwise deletion) that tend to work well in a very limited set of circumstances. For example, a recent review of published articles from the 2003 volumes of 23 educational and psychological journals suggested that deletion methods (cases with missing values are completely discarded from the data set, or are discarded on an analysis by analysis basis) are the predominant techniques used to handle missing data (Peugh & Enders, 2004). Intuition suggests that discarding cases with incomplete data may bias subsequent statistical analyses, because students with complete data may possess different characteristics than those with missing values (e.g., they may differ in their disability classifications, family structure, ability, peer relationships, etc.). Referring to deletion methods (i.e., listwise and pairwise deletion), a report by the APA (American Psychological Association) Task Force on Statistical Inference stated that these techniques are "among the worst methods available for practical applications" (Wilkinson & Task Force on Statistical Inference, 1999, p. 598). A number of recent empirical studies support this statement (e.g., Enders & Bandalos, 2001).

Two "modern" missing data methods, full information maximum likelihood estimation (FIML) and multiple imputation (MI), have received considerable attention in the methodological literature during the last 20 years, and are currently considered to be the "state of the art" (Schafer & Graham, 2002, p. 173) missing data techniques. These methods are advantageous because they require a less strict assumption about the missing data, and thus will provide unbiased parameter estimates in situations where traditional methods will not. Because these methods use all available data, they also tend to be more efficient (i.e., powerful) than traditional missing data handling techniques. Finally, FIML and MI are widely available in popular software packages, and should provide an attractive analysis option for researchers who face the difficult problem of attrition in special education studies.

It has been predicted that the routine implementation of FIML and MI will be one of the major changes in research methodology during the course of the next decade (Stephen G. West, cited in Azar, 2002). As such, the purpose of this chapter is to familiarize researchers with FIML and MI, and demonstrate the use of these techniques in the context of a special education study (Montague, Enders, & Castro, 2005). We begin with a brief overview of Rubin's (1976) missing data mechanisms, as this provides a theoretical foundation for comparing the performance of different missing data techniques. Next, we review several missing data methods, and discuss the assumptions associated with each. Brief descriptions of traditional techniques are given, but the primary focus is on FIML and MI, as these are the recommended procedures in the methodological literature. Finally, we demonstrate the use of FIML and MI using data from a longitudinal study of adolescents at risk for the development of emotional or behavioral disorders (Montague et al., 2005).

A brief description of the Montague et al. (2005) study is warranted at this point, as the concepts presented in this chapter will be explored in the context of this research scenario. The at risk students in this longitudinal study were originally identified by screening 628 kindergarten and first-grade students in two urban elementary schools using the Systematic Screening for Behavior Disorders (Walker & Severson, 1992). In this initial screening, 115 students were identified as low risk, 63 as moderate risk, and 28 as high risk. Of these, 113 students were located through the district database when they were in grades 7 and 8, and a cohort of 99 not at risk students were subsequently added ($n = 212$). A variety of measures focusing on school, family, and person–social variables are being administered twice yearly for a total of 10 data waves. One of the measures is the Children's Depression Inventory (CDI; Kovacs, 1992) consisting of 27 items, which is intended as a screening measure for identifying depressive symptomology. This measure will be used to illustrate the problems and solutions associated with missing data in a longitudinal study.

STATISTICAL THEORY

Rubin's (1976) seminal theoretical work provided a taxonomy of missing data mechanisms. In this case, the term "mechanism" is not intended to convey a causal relationship, but is a probabilistic explanation for how the missing values are related to variables in the data set. Taking a slightly different view, Rubin's taxonomy can be viewed as a set of assumptions that dictate the performance of a given missing data technique. As alluded to above, FIML and MI are theoretically advantageous because they require a less strict assumption about the missing data, and thus should provide accurate results in situations where traditional methods will not.

According to Rubin (1976), data are *missing at random* (MAR) if the probability of a missing value is (a) related to other measured variables, but (b) unrelated to the underlying values of the variable that are missing. Unfortunately, it is natural to interpret the phrase "missing at random" to mean that missing values are governed by a process resembling a coin toss. However, MAR actually means that missingness is related to measured variables *other than* the specific outcome variable of interest. In the context of the longitudinal study, MAR would hold if the probability of a missing depression score was related to the values of predictor variables or to CDI scores from previous assessments, but not to the severity of depression *at the particular assessment that is missing*. A number of plausible examples of MAR data can be generated from the Montague et al. (2005) study. For example, student mobility (which itself is likely related to socioeconomic status) is a pervasive problem when conducting studies in large urban school districts. For many students, attrition may result from transferring to a different school within the district. As long as the transfer had nothing to do with depression (or disruptive behaviors due to the presence of depressive symptoms), this situation would be described as MAR. As a second example, the state of Florida requires students to pass a statewide assessment (the Florida Comprehensive Achievement Test; FCAT) in order to graduate from high school. For some secondary students, it is possible that attrition is systematically related to FCAT scores, such that students who fail the test might be at risk for dropping out of school (and the study). Unfortunately, MAR is a condition that cannot be empirically verified from the data (doing so would require knowledge of the missing values). However, researchers involved in the Montague et al. (2005) study made exhaustive attempts to track and contact students who left the study, and the results of these follow-up interviews suggested that MAR was plausible in many situations. The results of these follow-up contacts are summarized in a subsequent section.

The *missing completely at random* (MCAR) mechanism is a special case of MAR, with the additional requirement that missingness is unrelated to the observed data. In this case, the observed scores can be viewed as a random sample of the hypothetically complete data set (MCAR is actually more closely aligned with the notion of a random coin toss). To illustrate, suppose that a small number of scores were missing because the paper copies of the assessments were inadvertently misplaced before the data could be entered.

Similarly, missing values might result due to any number of scheduling difficulties (e.g., a child's record was missing due to a family vacation or a doctor's appointment on the date of the assessment). Returning to the CDI study, one child had a number of missing assessments because he was killed in an unfortunate bicycle accident. In this case, it would be difficult to argue that death was systematically related to any measured variable in the data set, so this child's data could be viewed as MCAR.

The last of Rubin's (1976) missing data mechanisms is referred to as *missing not at random* (MNAR). Data are MNAR if the probability of missing data is systematically related to the values that are missing. For example, suppose that a number of CDI assessments were missing for a student who dropped out of school due to severe depression. In this case, the reason for the missing CDI scores is directly related to symptom severity, so these data would be described as MNAR. This situation was observed in the Montague et al.'s (2005) study when a school counselor informed members of the research team that a particular student had been hospitalized for "psychiatric reasons". Although it was not specifically stated that this student was being treated for clinical depression, it seems safe to conclude that this student's data were MNAR, given that psychiatric hospitalization often occurs when a student is at risk of hurting himself or others (i.e., clinically depressive symptomology).

Rubin's (1976) missing data mechanisms are important because they dictate the situations in which missing data techniques will provide optimal performance. For example, most traditional missing data procedures (e.g., deletion methods) require the MCAR assumption. This is a restrictive condition that some methodologists feel is rarely met in practice (Muthén, Kaplan, & Hollis, 1987; Raghunathan, 2004). In contrast, FIML and MI require the more relaxed MAR assumption, and should produce unbiased statistical estimates when data are either MCAR or MAR. Additionally, FIML and MI will generally be more powerful, even when MCAR does hold (substantially more powerful, in some cases; Enders & Bandalos, 2001). Despite these advantages, MNAR may be the most plausible explanation for why values are missing in some situations. In this case, FIML and MI will yield biased parameter estimates (as will traditional techniques). MNAR missing data methods have been proposed in the literature, but these methods are difficult to implement due to a lack of software, and are prone to substantial bias when the user does not correctly specify a model for the missing data. Because FIML and MI currently represent the "practical state of the art" (Schafer & Graham, 2002, p. 173), we chose to limit our discussion to these techniques.

One final point needs to be made about the MAR assumption required by FIML and MI. MAR (or any missing data mechanism, for that matter) is not an inherent characteristic of a data set, and may or may not hold for different analyses performed on the same database. Even when the true "cause" of the missing data is captured in the data (e.g., dropout is systematically related to poor test performance), MAR will only hold if the variable related to missingness is included in the analysis model. For example, consider an earlier example where we raised the possibility that attrition was systematically related to low scores on the Florida state assessment, the FCAT. Even if such a relationship existed in our data, MAR would only hold if FCAT scores were somehow included in the longitudinal analysis. However, including FCAT as a predictor variable fundamentally alters our substantive research question, as it was not of interest to estimate CDI growth, conditional on achievement test performance.

Fortunately, there are established methods for incorporating information from auxiliary variables into a missing data analysis (Graham, 2003). An auxiliary variable can be defined as a variable that is unrelated to one's substantive hypotheses (i.e., would not appear in the analysis model, had the data been complete), but may be (a) a potential cause or correlate of missingness, or (b) a correlate of the variable that contains missing values. Throughout this chapter we give special emphasis to an "inclusive" analysis strategy (Collins, Schafer, & Kam, 2001) that incorporates auxiliary variables into the statistical analysis, as this strategy can make the MAR assumption more plausible, but can also improve the accuracy of the results obtained from a missing data analysis.

HOW PLAUSIBLE IS THE MAR ASSUMPTION?

Unfortunately, only the MCAR assumption can be empirically tested from the data (Little, 1988), as a test of MAR or MNAR would require knowledge of the missing values. This means that researchers must generally proceed with a FIML or MI analysis by adopting the important, albeit untestable, assumption that the data are MAR. However, evidence to support the MAR assumption can be established through follow-up interviews. For example, Graham, Hofer, Donaldson, MacKinnon, and Schafer (1997) described a longitudinal study of substance use where attrition was most frequently linked with student mobility rather than substance use itself. In a similar vein, members of Montague et al. (2005) research team made exhaustive attempts to track and contact students who left the study, and the results of these follow-up interviews suggested that the reason for attrition could likely be characterized as MAR in many cases. Although some caution is warranted in generalizing our results to other contexts, we briefly describe some of the reasons why data were missing in the Montague et al. (2005) study. In doing so, we hope to underscore the importance of carefully considering and *empirically examining* the potential reasons for missingness using information gathered from follow-up contacts.

As noted previously, student mobility is a pervasive issue when considering attrition, particular in a population of students from a large urban school district. In some cases, relocation to a different school may result from behavioral problems related to the outcome of interest. However, we believe that, in many cases, mobility-related attrition is largely a function of SES. For example, our research team encountered difficulties contacting some students because their families did not have working telephone numbers (presumably due to financial hardships). In the absence of reliable demographic data, a proxy such as free or reduced lunch status may serve as a useful auxiliary variable in the missing data analysis. In our study, this variable was of limited utility because the vast majority of students qualified for lunch assistance (i.e., the variable had little variability, and thus could not bear any relationship with missingness). Related to the SES issue, we found that some students were not accessible during school hours because they were employed during the day. Situations such as this are likely more common as students get older and opt to participate in community work study programs in lieu of attending classes in a traditional academic setting.

In addition to SES-related missingness, we encountered a number of other situations where attrition was arguably unrelated to the missing CDI scores. For example, a number of female students were unable to be reached because they were at home on maternity leave or were on medically ordered bed rest. A number of students in our study were also missing assessments because they were involved in the juvenile justice system (e.g., they were incarcerated or were attending court hearings on the assessment date). Unfortunately, there was a significant communication gap between the school system and the juvenile justice system, making it difficult to continue collecting data from students following their incarceration. In one situation, the third wave of data collection actually took place in the juvenile detention hall with a bodyguard present. This student subsequently transferred to a different school following his release, but refused to participate in subsequent data collection waves, despite repeated attempts.

The academic characteristics of a student also played an important role in our ability to obtain complete data. For example, finishing an assessment battery is often difficult for students who are categorized as having learning disabilities, emotional or behavioral disorders, and those who exhibit academic or behavioral characteristics that are consistent with receiving special education services. These students frequently become frustrated, fatigued, and lack the motivation to complete the assessment tasks within the specified time frame (approximately 1.5 h). Such behavior is not atypical when students with learning, emotional, or behavioral difficulties are asked to complete academic tasks that require patience and concentration. Obtaining complete data from these students is difficult, and frequently requires multiple visits to the school in order to complete a single assessment battery. In many cases it is only possible to collect partial data from these students (e.g., BASC scores may be complete, but CDI scores are missing). The fact that these students also have a higher likelihood of suspensions and are frequently absent only serves to exacerbate missing data problems.

We found relatively few cases that could be unequivocally characterized as MNAR. As described above, one student was missing a number of assessments because he had been hospitalized for "psychiatric reasons" – presumably, this meant that he was experiencing clinically depressive symptomology. Although it was difficult to find frequent examples of missing records that were directly related to depressive symptoms, it would be incorrect to conclude that our entire database satisfies MAR. For example, suppose that we were interested in analyzing the development of aggressive or antisocial behavior. In this case, the missing data for those students who were incarcerated or were involved with the juvenile justice system might be classified as MNAR. The examples given in this section also illustrate the point that missing values should not be viewed as falling into one of three mutually exclusive mechanisms. Rather, it is more likely the case that some missing records are MCAR, while others are MAR or MNAR.

SOME TRADITIONAL MISSING DATA TECHNIQUES

A brief overview of some traditional missing data techniques is warranted before discussing FIML and MI. Literally dozens of ad hoc missing data techniques have been proposed in the literature, many dating back several decades. Space limitations preclude an exhaustive review of these methods, so we chose to discuss only those methods that appear routinely in published education studies (Peugh & Enders, 2004). Readers are encouraged to consult Little and Rubin (2002) for a comprehensive treatment of these techniques.

Listwise Deletion

Listwise Deletion (LW, also referred to as *complete case analysis*) removes incomplete records from the database prior to analysis. This method is appealing due to its simplicity, because subsequent statistical analyses are performed using only the complete portion of the data. However, removing cases can result in a dramatic reduction in sample size, and thus power. More importantly, LW typically requires the MCAR assumption, and may produce substantial bias when this strict condition does not hold (e.g., Enders & Bandalos, 2001).

Pairwise Deletion

Pairwise Deletion (PW) attempts to retain as much of the data as possible, and deletes cases on an analysis by analysis basis. For example, suppose it was of interest to compute the covariance matrix for a set of variables. Each element in the PW covariance matrix would be computed using the cases with complete data on a given variable (for variances) or variable pair (for covariances). A number of problems with PW have been noted in the literature (e.g., no single *N* is applicable, the possibility of non-positive definite matrices), so this procedure is generally not recommended (Wilkinson & Task Force on Statistical Inference, 1999). Like LW, the MCAR assumption is a problematic aspect of PW.

Arithmetic Mean Imputation

Arithmetic Mean Imputation (AMI) replaces missing values with the arithmetic mean of the complete cases. Although AMI produces accurate mean estimates under MCAR, estimates of variation and covariation (and thus any statistics related to covariation) are substantially negatively biased; missing values are imputed at the center of the score distribution, which dramatically reduces the variation, and thus measures of association. The bias due to mean imputation has been documented in a number of empirical studies (e.g., Enders, 2001; Wothke, 2000), and Little and Rubin (2002) stated that AMI "cannot be recommended" (p. 62).

Regression Imputation

Regression imputation (RI) replaces missing values with predicted scores from a linear regression equation. Because the imputed values fall directly on a regression line (or surface), the imputed data will understate the amount of variation present in the hypothetically complete data, resulting in negatively biased variance and covariance estimates. Stochastic regression imputation (SRI) attempts to restore variation to the imputed data by adding a randomly sampled residual term to each imputed value. Among the traditional missing data techniques described here, SRI is probably preferred, because it performs well under the less stringent MAR assumption.

Having provided brief descriptions of some common traditional missing data techniques, we now provide a more detailed overview of the so-called "modern" missing data techniques, FIML and MI. These procedures are theoretically advantageous because they require a less strict assumption about the missing data (MAR), which means that FIML and MI should provide accurate parameter estimates in situations where traditional methods will not – as noted previously, biased parameter estimates would be expected with MNAR data.

FULL INFORMATION MAXIMUM LIKELIHOOD

Maximum likelihood (ML) estimation is routinely used to estimate complex statistical models (e.g., structural equation models, SEM), and is well-suited for missing data problems as well. The basic goal of ML estimation is to identify the population parameters that are most likely to have produced the sample data. Conceptually, different values for the unknown parameters are "auditioned" using iterative algorithms, and values are chosen that maximize the log likelihood. Accessible introductions to ML estimation can be found in Enders (2005) and Eliason (1993).

The fit of a set of parameter values to the raw data is quantified by the log likelihood. In the missing data context, a log likelihood value is computed for each case using all available data for that case (in the CDI growth curve analysis, a child with only one complete data point would contribute to the analysis). Assuming a multivariate normal distribution, the log likelihood value for case i is

$$\log L_i = K_i - \frac{1}{2}\log|{\textstyle\sum_i}| - \frac{1}{2}\left[(\mathbf{x}_i - \boldsymbol{\mu}_i)' {\textstyle\sum_i^{-1}} (\mathbf{x}_i - \boldsymbol{\mu}_i)\right] \tag{1}$$

Collectively, the terms inside the brackets are referred to as Mahalanobis distance, and are comprised of the raw data vector, $\mathbf{x}_i$, the estimated mean vector, $\boldsymbol{\mu}_i$, and the estimated covariance matrix, $\boldsymbol{\Sigma}_i$ (K_i is a scaling factor that depends on the number of complete data points for case *i*, and can be ignored during estimation). The important point about Eq. (1) is that the raw data vector need not be complete – the size and contents of the parameter arrays are adjusted, such that Mahalanobis distance is computed using only the variables and parameters for which case *i* has complete data. Consistent with complete-data ML, parameter estimates (e.g., $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) are sought that maximize the sample log likelihood, which is the sum of Eq. (1) over the *N* cases.

To illustrate, suppose it is of interest to estimate the covariance matrix and means for the first four waves of CDI data. Furthermore, suppose that a subset of cases is missing CDI scores from the second and third assessment. The contribution to the log likelihood for these cases would be computed as follows:

$$\log L_i = K_i - \frac{1}{2}\log\begin{vmatrix}\sigma_{11} & \sigma_{14}\\ \sigma_{41} & \sigma_{44}\end{vmatrix} - \frac{1}{2}\left(\begin{bmatrix}cdi_1\\ cdi_4\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_4\end{bmatrix}\right)'\begin{bmatrix}\sigma_{11} & \sigma_{14}\\ \sigma_{41} & \sigma_{44}\end{bmatrix}^{-1}$$
$$\times\left(\begin{bmatrix}cdi_1\\ cdi_4\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_4\end{bmatrix}\right)$$

In a similar vein, the log likelihood for cases that are missing the fourth assessment is shown below.

$$\log L_i = K_i - \frac{1}{2}\log\begin{vmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{vmatrix} - \frac{1}{2}\left(\begin{bmatrix}cdi_1\\ cdi_2\\ cdi_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}\right)'$$
$$\begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{bmatrix}^{-1}\left(\begin{bmatrix}cdi_1\\ cdi_2\\ cdi_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}\right)$$

In both of these examples, notice that the rows and columns corresponding to the missing values are simply removed, and the fit of the raw data to the parameters is based only on the observed data. Note also that the parameter values themselves do not change from one pattern of missingness to the next (e.g., the estimate of μ_1 is identical in both patterns shown above).

It is important to note that the derivation of Eq. (1) relies explicitly on the multivariate normality assumption. Although parameter estimates tend to be accurate when data are nonnormal, standard errors will be too low, resulting in increased Type I error rates (Enders, 2001). Readers with previous exposure to SEM may be familiar with so-called "robust" fit statistics and standard errors. These robust statistics have recently been developed for missing data (Yuan & Bentler, 2000), and are available in some commercial SEM software packages (EQS 6.0, M*plus* 3.0).

It may not be obvious from Eq. (1), but FIML actually "borrows" information from the observed data when estimating parameters associated with variables that have missing values. Although the missing values themselves are not imputed, the inclusion of partially complete cases does imply probable values for the missing data, and does so via the correlations among the variables. To illustrate, a small artificial data set consisting of 10 CDI scores from three assessments (cdi_1, cdi_2, cdi_3) is given in Table 1. Missing values were created on cdi_2 and cdi_3 according to an MAR mechanism, such that missing values were isolated to the four cases with the highest cdi_1 scores (mimicking a situation where children who are highly depressed at the onset of the study are more likely to be missing). Descriptive statistics and correlations were obtained using the EM algorithm (an iterative algorithm that yields FIML estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$), as implemented in the SPSS Missing Values Analysis (MVA) procedure. ML parameter estimates are given in Table 2, and results from a LW analysis are also presented for comparison purposes.

Although this demonstration was clearly "rigged" in favor of FIML (according to theory, ML should perform better than LW under MAR), the results in Table 2 do demonstrate several interesting points. First, consider the LW estimates, all of which were severely biased (e.g., the LW correlation

Table 1. Artificial CDI Data with MAR Mechanism.

Complete Data			Missing Data		
CDI1	CDI2	CDI3	CDI1	CDI2	CDI3
39	47	50	39	?	50
30	32	24	30	?	24
30	32	40	30	32	?
29	27	29	29	27	?
27	31	35	27	31	35
27	31	26	27	31	26
27	24	23	27	24	23
26	24	27	26	24	27
24	29	36	24	29	36
21	27	37	21	27	37

Table 2. Descriptive Statistics from Artificial CDI Data.

Estimate	Variable	Means	Correlations		
			CDI1	*CDI2*	*CDI3*
Complete data	*CDI1*	28.00	4.74		
	CDI2	30.40	0.83	6.57	
	CDI3	32.70	0.46	0.75	8.49
FIML	*CDI1*	28.00	4.74		
	CDI2	29.49	0.75	5.30	
	CDI3	32.01	0.42	0.77	8.59
LW ($n = 6$)	*CDI1*	25.33	2.42		
	CDI2	27.67	0.09	3.20	
	CDI3	30.67	−0.69	0.45	6.02

Note: Diagonal elements of correlation matrices contain standard deviations. FIML = full information maximum likelihood. LW = listwise deletion.

between cdi_1 and cdi_3 was $r = -0.69$, as compared to $r = 0.46$ for the complete data). Recall that missing values were systematically related to high cdi_1 scores. Because the three CDI variables had strong positive correlations (consistent with a longitudinal study), the listwise removal of incomplete cases served to truncate the high end of the CDI score distribution, producing biased estimates – this is clearly evident in the LW means, all of which are too low. In contrast to LW, FIML estimates were relatively accurate. The inclusion of partially observed data (e.g., case 1) essentially "steered" the estimation algorithm toward a different, and more accurate, set of parameter estimates than would be obtained had the incomplete cases been discarded. That is, the inclusion of the high scoring (albeit incomplete) cases on cdi_1 provided information that improved the estimates of the distributional properties of cdi_2 and cdi_3.

Incorporating Auxiliary Variables

As noted previously, MAR only holds if the variable related to missingness is included in the analysis model. For example, we previously raised the possibility that attrition may be related to FACT achievement scores for some students, given that a satisfactory score on the assessment is necessary for graduation. This potential relationship needs to be captured in the analysis in order for MAR to hold, but the substantive research question was related to CDI growth, and achievement test scores were of no interest.

Fortunately, there are established methods for incorporating auxiliary variables into a FIML analysis (Graham, 2003), and these methods are relatively straightforward to implement using any of the commercial SEM software packages.

If the only analytic interest was to estimate correlations among a set of variables, the inclusion of auxiliary variables is straightforward – simply obtain ML estimates of the correlations (e.g., using the EM algorithm implemented in the SPSS MVA procedure), but do so using a superset of the variables that are of substantive interest (i.e., estimate the correlations of interest, but also estimate the correlations between the substantive and auxiliary variables). In the more general case, the linear model of interest (e.g., regression, growth curve analysis, etc.) can be estimated using SEM software, and the auxiliary variables can be incorporated using the ‘‘saturated correlates’’ approach proposed by Graham (2003).

Graham's (2003) approach incorporates auxiliary variables into the FIML analysis, but does so in a way that does not alter the meaning of the substantive parameters, because the extraneous variables are not included as additional predictors in the model. Three simple rules must be followed when incorporating auxiliary variables into a FIML analysis: an auxiliary variable must be (a) correlated with all other auxiliary variables, (b) correlated with all observed predictor variables, and (c) correlated with the residual term from any observed criterion variable. It is important to note that these rules do not apply to latent variables; under no situation should an auxiliary variable be correlated with a latent factor. In addition to the examples given in Graham's manuscript, examples of auxiliary variables in the context of SEM and multiple regression can be found in Enders (2006) and Peugh and Enders (2004), respectively. We demonstrate the use of auxiliary variables in the growth model analysis presented in the final section of this chapter.

MULTIPLE IMPUTATION

The goal of MI imputation phase is to create multiple copies of the data (e.g., $m = 10$), each of which is imputed with slightly *different* estimates of the missing values. In the subsequent analysis phase, the desired statistical model (e.g., the CDI growth model) is fit to each of the m complete data sets, and the resulting parameter estimates and standard errors are combined into a single estimate using arithmetic rules given by Rubin (1987). A key contrast with FIML is that the imputation and analysis phases are distinct – recall that FIML does not impute missing values, but missing data handling is integrated into the process of estimating model parameters. In fact, it is quite possible for a number of different analyses to draw upon the same set of imputations, and a single set of well-planned imputations can be shared among a number of researchers.

Although the process of analyzing multiple data sets and combining parameter estimates sounds tedious, a number of software packages now have routines designed to automate the analysis of multiply imputed data (e.g., SAS, HLM, M*plus*) – we demonstrate such a procedure using M*plus* later in this chapter. A number of MI algorithms have been proposed in the literature (e.g., see Allison, 2000), but we focus on Schafer's (1997) data augmentation (DA) procedure, as this is arguably the most popular, and is readily available in software packages (e.g., SAS, NORM).

Like FIML, MI also relies on the multivariate normality assumption. Computer simulation results from Graham and Schafer (1999) suggested that MI performs reasonably well, even when normality is violated. Furthermore, Schafer (1997) suggested that nominal and ordinal variables can be used in many cases (nominal variables must be represented by a set of dummy variables). MI software packages such as SAS and NORM offer the user a number of normalizing transformations that can be implemented prior to the imputation phase, and variables can subsequently be restored to their original metrics prior to analysis.

Imputation Phase

Schafer's (1997) DA algorithm iteratively cycles between two steps, the *imputation* (I) and *posterior* (P) steps. The DA algorithm begins with an initial estimate of the mean vector and covariance matrix, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. The basic idea of the I step is to impute missing values with predicted scores from a set of regression equations that are constructed from $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Because these predicted scores fall directly on a regression surface, randomly sampled residuals are added to each imputed value in order to restore variation to the filled-in data. After the missing values have been imputed, an updated estimate of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is computed from the complete data set.

As noted above, each of the imputed data sets is filled in with different estimates of the missing values. This is accomplished in the P step, in what amounts to adding random perturbations to the regression equations used to generate the imputed values. More specifically, new values for the elements in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are randomly sampled from a distribution of parameter estimates (called a *posterior distribution*) that is conditional on the filled in data at the previous I step. These updated estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are carried forward to the next I step, where new imputed values are generated from regression equations that differ slightly from those in the previous I step. This two-step procedure is repeated a large number of times (e.g., 1,000), and imputed data sets are saved at specific intervals in the sequence (e.g., after every 100th I step).

The ultimate goal of the DA algorithm is to create *m* imputed data sets, such that the filled-in values in any single data set are independent of the imputed values in the remaining $m-1$ data sets. However, the DA algorithm produces imputations that are correlated from one I step to the next. In order to achieve independent imputations, it is necessary that a number of DA iterations (I and P steps) separate the imputed files that are ultimately saved for further analysis. For example, 500 *between-imputation iterations* were specified in the subsequent CDI analysis, meaning that an imputed data set was saved at every 500th I step (the data files created at the intermediate I steps are simply discarded).

For different reasons, it is also necessary to allow a number of DA iterations to lapse before saving the *first* imputed data set (these preliminary cycles are sometimes called *burn-in iterations*). For example, a total of 1,000 burn-in iterations were specified in the CDI example below. The optimal number of between-imputation and burn-in iterations will vary across data sets, and incorrectly specifying these values can impact on the quality of the imputed values. In order to choose the correct values, it is necessary to assess the convergence (i.e., stability) of the sampled values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that are drawn at the P step. This process is aided by graphical displays (time series and autocorrelation function plots) that allow the user to assess the magnitude and duration of the correlation that exists across successive DA iterations. Schafer (1997) provided a detailed discussion of graphical techniques, and Schafer and Olsen (1998) provide an accessible example using Schafer's (1999) NORM freeware package.

Analysis Phase

The imputation phase produces *m* complete data sets, so standard statistical software can be used for all subsequent analyses. Having created the imputed data sets, the analysis phase consists of fitting the desired statistical model (e.g., the CDI growth curve model) to each of the *m* data sets. The *m* sets of parameter estimates and standard errors are subsequently combined into a single inference, following rules given by Rubin (1987). Parameter values are combined into a single point estimate by taking the arithmetic average of the parameter across the m analyses, as follows

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i \qquad (2)$$

where $\hat{Q}_i$ is the parameter estimate from the ith data set, and m the number of imputations.

Standard errors are combined in a similar fashion, but require the calculation of two components. The *within-imputation variance* is computed by taking the arithmetic average of the m squared standard errors, as follows

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} \hat{U}_i \qquad (3)$$

where $\hat{U}_i$ is the squared standard error from the ith data set. The *between-imputation variance* is the variance of the parameter estimate itself across the m imputations, or

$$B = \frac{1}{m}\sum_{i=1}^{m} \left(\hat{Q}_i - \bar{Q}\right)^2 \qquad (4)$$

Finally, the MI standard error combines both the within- and between-imputation variance, as shown below

$$\text{S.E.} = \sqrt{\bar{U} + \left(1 + \frac{1}{m}\right)B} \qquad (5)$$

Having collapsed the results from the m analyses into a single set of point estimates and standard errors, parameter estimates can be tested for statistical significance using a t statistic, computed as the ratio of the MI parameter estimate by its associated standard error (i.e.,$\bar{Q}$/S.E.). Simultaneous tests of multiple parameters (i.e., akin to an omnibus F test) can also be obtained using multivariate inference outlined by Li, Raghunathan, and Rubin (1991). Software packages that analyze multiply imputed data sets may offer the user different options for computing the degrees of freedom for the t statistic. Whenever possible, the user should request the degrees of freedom outlined by Barnard and Rubin (1999), as these are generally considered to be superior.

Incorporating Auxiliary Variables

One compelling advantage of MI is the ease with which auxiliary variables can be incorporated. Auxiliary variables are simply added as predictor variables during the imputation phase, and can be ignored during all subsequent analyses (the filled-in values are already conditioned on the extra variables). Consistent with the previous discussion, auxiliary variables should be chosen that are related to the variable being imputed, and potentially related to missingness on that variable. In addition to auxiliary variables, the imputation model should include all effects that are of interest in the subsequent analyses (i.e., the imputation model should be at least as general as the analysis model). For example, if it was of interest to perform an analysis that included an interaction, an interaction (i.e., product) term should be computed and included in the imputation model – failure to do so could attenuate the interaction effect in the subsequent analysis. Although the imputation and analysis phases are distinct and need not be performed by the same user, this point underscores the need to carefully consider future analyses when choosing variables for the imputation phase. Schafer (1997, p. 139) provides a detailed discussion of this process.

ANALYSIS EXAMPLE

Having outlined FIML and MI, we now present an example analysis using seven waves of CDI data from the Montague et al. (2005) longitudinal study. The development of CDI scores was examined using a linear growth model. A brief description of growth modeling is given here, but interested readers can obtain more details from a number of different sources in the literature (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003).

The linear growth model for individual i is given by

$$Y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + e_{ti} \tag{6}$$

where Y_{ti} is the outcome variable (i.e., CDI) score for individual i at assessment t, π_{0i} an intercept (e.g., estimated initial status), π_{1i} the growth rate (e.g., change per year), and e_{ti} the level-1 residual that captures the deviation between an individual's observed data and their idealized linear growth trajectory. The a_{ti} term in Eq. (6) is a variable that captures the timing of the repeated measures for person i. In the current example, it was of interest to examine CDI growth as a function of age, so a_{ti} represented a student's age at assessment t. Furthermore, student age was centered at a value of 13 (i.e., $a_{ti} = \text{age}_{ti}-13$) in order to yield a more straightforward interpretation of the intercept, π_{0i}. Under this centering scheme, the intercept is defined as an individual's estimated CDI score at age 13 (i.e., the estimated CDI value when a_{ti} equals 0), and π_{1i} represents the yearly change in CDI. The so-called level-1 equation given above represents the linear growth for an individual. Conceptually, the growth curve analysis aggregates the values of π_{0i} and π_{1i} to produce a mean intercept and slope, but also yields variance estimates of π_{0i} and π_{1i} that quantify individual variation in initial status and growth rates. All growth curve analyses were performed using M*plus* 3.13 (Muthén & Muthén, 2004), a structural equation modeling software package. The use of M*plus* was advantageous because it allows the timing of the assessment schedules (i.e., the a_{ti} values) to vary across individuals, much like the multilevel growth model formulation. As seen in the M*plus* syntax given in the appendices, person-specific assessment schedules are read in as data using the TIMESCORES option.

To illustrate the use of auxiliary variables, we identified a small set of variables that might potentially be related to missingness – the choice of auxiliary variables was based on our analysis of the follow-up data discussed earlier. The auxiliary variables used in the growth curve analysis included gender, special education status (a binary dummy variable), four BASC teacher subscales (BSI, school conduct problems, attention problems, and learning difficulties), the number of unexcused absences during the prior school year, the number of days spent on suspension during the last school year, and FCAT math and reading scores. To reduce the number of auxiliary variables, the BASC and FCAT subscales were submitted to separate principle components analyses, and linear combinations of these variable sets (i.e., factor scores) were used in the subsequent analyses. It is most certainly the case that additional auxiliary variables could be identified, but we chose to limit the number of variables in the present demonstration.

Before proceeding to the analyses, we began by testing whether the MCAR assumption was plausible using the multivariate test proposed by Little (1988). Little's (1988) MCAR test is available in the SPSS MVA procedure, and a custom SAS macro program has been made available for download at www.asu.edu/clas/psych/people/faculty/enders.htm. The null hypothesis for Little's test states that missingness is unrelated to the variables in the data set. Using the seven waves of CDI scores and the auxiliary variables as input data, we rejected the MCAR hypothesis, $\chi^2(391) = 546.71$, $p<0.001$. Again, we cannot empirically verify that MAR holds, but our follow-up data suggest that this assumption is plausible for many students in our database.

FIML Analysis

Auxiliary variables were incorporated into the FIML analysis using Graham's (2003) saturated correlates approach. A graphical depiction of the growth model is given in Fig. 1, and the corresponding M*plus* syntax is given in Appendix A. Note that the path diagram in the figure does not exactly correspond with our analysis model; to reduce clutter in the graphic, only four repeated measures variables and two auxiliary variables are shown.

A number of points should be made about the M*plus* program in Appendix A. First, the metrics of the auxiliary variables were quite different from one another, which initially resulted in convergence problems. Following standard recommendations in the literature (e.g., Muthén &

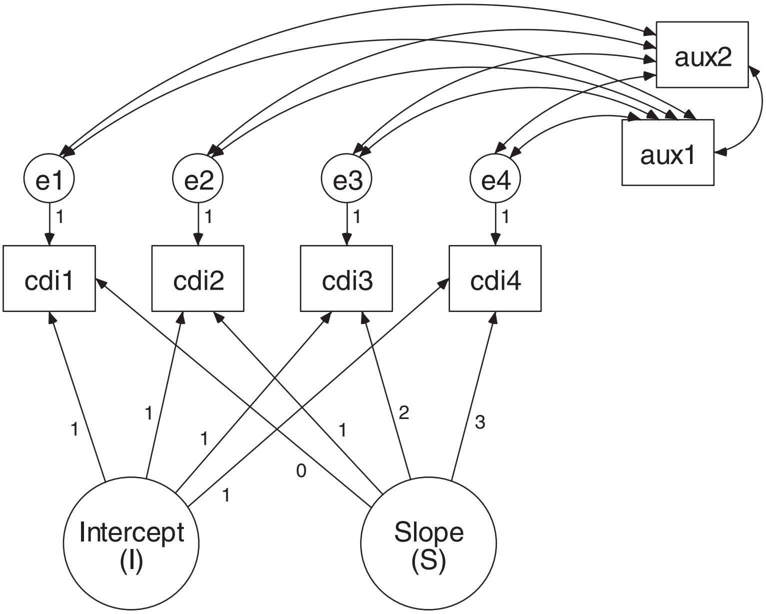

Fig. 1. Path Diagram Representation of the Latent Growth Curve Model. Note that the Auxiliary Variables are Correlated with each other, and with the Residual Terms of each Repeated Measure.

Muthén, 2004), the variances of the variables were made more similar by multiplying or dividing scores by a constant value (e.g., the gender dummy variable was multiplied by 10 in order to increase the value of its variance). As seen in the appendix, the variables were rescaled using the M*plus* DEFINE command. Next, the magnitude of the residual (i.e., level-1) variance values were held constant across data collection waves. Although this is not necessary when using SEM software, invoking this constraint produces a model that is equivalent to the multilevel formulation of the growth curve model (the multilevel model assumes that the level-1 residual variance is constant). This equality constraint was imposed by listing a numeric value in parentheses after naming the residual variances (e.g., this shows up as (1) in the M*plus* program). Finally – and most importantly – the auxiliary variables are incorporated into the model using only two additional lines of code. As seen in the appendix, a single line of code is used to specify the correlations between the auxiliary variables and the residual terms (i.e., cdi1–cdi7 with male–fcat;), and a second line is used to specify correlations among the auxiliary variables (i.e., male–fcat with male–fcat;).

MI Analysis

The SAS MI procedure was used to create 10 imputed data sets, the syntax for which is given in Appendix B. Note that the MI procedure was initially implemented in version 8.1 of SAS, and is not available in earlier releases of SAS. For those who do not have access to SAS, Schafer's (1999) NORM software can be freely downloaded at www.stat.psu.edu/~jls/misoftwa.html. NORM offers a convenient point and click interface, and offers a similar range of options as SAS. Readers who are interested in more details about the use of NORM can consult Schafer and Olsen (1998) and Peugh and Enders (2004).

The number of imputed data sets is specified using the NIMPUTE option on the PROC MI command line. As seen in the appendix, the predictor variables used during the imputation phase are listed on the VAR subcommand line. In this case, the seven waves of CDI scores are listed along with the auxiliary variables and the age of the child at each assessment (as described previously, age is the temporal predictor variable in the level-1 growth model). Consistent with Schafer's (1997, p. 148) recommendations, the imputed CDI values were rounded to the nearest integer, as this was consistent with the original metric of the CDI. The ROUND option was used to specify the rounding precision of each variable listed on the VAR line, and a value of 1 means that a variable is to be rounded to the nearest integer (a period denotes no rounding). Note that the ordering of values listed after the ROUND option corresponds to the variable ordering on the VAR line. A minimum value for the CDI scale scores (according to the CDI manual, 34 is the lowest possible *T* score value) was given using the MINIMUM keyword; the MAXIMUM keyword also could have been used to set an upper bound on the imputed values, but none of the imputed CDI scores exceeded a value of 100. In cases where an imputed value falls outside the specified range, SAS simply generates a new imputed value. Like the ROUND option, minimum and maximum values are specified for each variable using the ordering of the variables in the VAR command.

Although the SAS syntax does not reflect this, preliminary analyses were conducted in order to determine the number of burn-in and between-imputation iterations. Graphical displays (time series and autocorrelation function plots) suggested that the DA algorithm converged relatively slowly (i.e., required a large number of iterations to stabilize), so we opted for a conservative approach, and specified a total of 1,000 burn-in iterations (i.e., the first imputed data set was saved after 1,000 I steps had lapsed) and 500 between-imputation iterations (i.e., additional data sets were saved every 500th iteration thereafter). Based on the graphical displays, we judged these to be *very* conservative values. However, specifying a large number of burn-in and between-imputation iterations was not problematic, because the entire imputation phase took only a few seconds on a modern microcomputer. These two options were specified in the SAS syntax using the NBITER and NITER keywords, respectively. When the imputation process is complete, the multiply imputed data files are stacked in a single file (specified by the OUT option on the PROC MI command line), and each imputed data file is indexed by variable named *_Imputation_* that ranges between 1 and *m*. Although it is not shown in the appendices, the imputed files were saved as 10 separate text files for later analysis using M*plus*.

SAS is a particularly convenient platform for MI because the multiply imputed data sets can be generated and analyzed within the same program (e.g., the growth curve analysis could have been performed using PROC MIXED). The MIANALYZE procedure can subsequently be used to combine the parameter estimates using Rubin's (1987) rules. However, we chose to analyze the multiply imputed data sets using M*plus*; the syntax for the analysis phase is given in Appendix C. M*plus* completely automates the process of analyzing and combining parameter estimates, so this seemingly tedious procedure becomes virtually transparent to the user.

Only two additions to the M*plus* code are needed when analyzing multiply imputed data sets. First, the TYPE = IMPUTATION subcommand is used to specify the input of multiply imputed data sets. It is normally the case that the FILE command is used to specify the input data set for the analysis. In this case, the FILE subcommand is used to specify a file containing the names of the imputed data sets. For example, the input data file (named cdifiles.txt) contains 10 rows, and each row lists the name of one of the imputed data sets (e.g., cdi1.dat, cdi2.dat, …, cdi10.dat). Upon executing the program file, M*plus* automatically fits the growth curve model to each of the imputed data sets, and combines the parameter estimates and standard errors according to the arithmetic rules outlined earlier. Again, note that the analysis model does not include the auxiliary variables, as these variables have already been accounted for during the imputation phase.

Analysis Results

Selected parameter estimates from the growth curve analyses are given in Table 3. As seen in the table, the FIML and MI parameter estimates were quite similar. The average initial CDI score was estimated at 48.18 using FIML, and depression scores declined by approximately 1.30 points per year, on average; these estimates are compared to values of 48.54 and −1.35 obtained from MI. The MI estimates of the intercept and slope variation were somewhat smaller in magnitude than those of FIML, but both sets of values suggested that substantial individual differences existed in CDI scores at age

Table 3. Selected Estimates from the Growth Curve Analysis.

Estimate	Technique	
	FIML	MI
Intercept mean	48.18	48.54
	(1.01)	(0.99)
Slope mean	−1.30	−1.35
	(0.31)	(0.30)
Intercept variance	127.56	96.98
	(35.24)	(24.34)
Slope variance	8.05	4.69
	(2.90)	(2.10)

Note: Values in parentheses are standard errors. FIML = full information maximum likelihood, MI = multiple imputation.

13, and in the rate of CDI growth over time. The close correspondence of the FIML and MI results are not surprising; Collins et al. (2001) noted that FIML and MI parameter estimates will generally be very similar, particularly when the same auxiliary variables are incorporated into the analysis.

DISCUSSION

The primary goal of this chapter was to introduce special education researchers to two "modern" methods of handling missing data, FIML and MI. These methods are appealing because they require a less strict assumption about the missing data (MAR), and are currently considered to be the "practical state of the art" (Schafer & Graham, 2002, p. 173) in the methodological literature.

After seeing the similarity between FIML and MI parameter estimates in Table 3, the reader may have surmised that there is no reason to prefer one technique over the other. In general, this is true. Both methods require identical assumptions (MAR and multivariate normality), and will generally produce very similar parameter estimates, given identical sets of variables (Collins et al., 2001). FIML may be the method of choice for researchers with previous exposure to SEM, given its widespread availability in commercial software packages. Although the inclusion of auxiliary variables into a FIML analysis is slightly awkward, this can be accomplished with little additional effort. Perhaps the biggest drawback of FIML is the availability of an estimation routine. The number of models that can be estimated using FIML has grown substantially in recent years, but there are still many analyses that cannot be performed using existing SEM software.

The availability of an estimation routine is not a drawback for MI, because the data analysis phase uses complete data sets. This means that virtually any analysis can be performed following the imputation phase, and parameter estimates can be combined using Rubin's (1987) arithmetic rules. The ease with which auxiliary variables can be incorporated into the MI imputation phase is also a benefit, as these variables can be ignored in all subsequent analyses. The primary drawback with MI is its complexity. Relative to FIML, MI is arguably more labor intensive, and requires more sophistication on the part of the user.

In the end, choosing between FIML and MI is probably a matter of personal preference and convenience. Empirical research and statistical theory suggest that both approaches offer a substantial improvement over the traditional missing data techniques that are still ubiquitous in the education literature. Of course, FIML and MI are not without their own difficulties, and are prone to bias when the MAR assumption is not met. Clearly, the best option is to implement rigorous data collection procedures that avoid the problem of missing data altogether. Unfortunately, the challenges associated with special education and at risk populations make such an idealistic statement a practical impossibility. For now, FIML and MI are the "state of the art" missing data techniques, and we hope that researchers begin to implement these methods with increased regularity.

REFERENCES

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research*, *28*, 301–309.

Azar, B. (2002). Finding a solution for missing data. *Monitor on Psychology*, *33*, 70.

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 948–955.

Blackorby, J., & Wagner, M. (1996). Longitudinal postschool outcomes of youth with disabilities: Findings from the National Longitudinal Transition Study. *Exceptional Children*, *62*, 399–413.

Collins, L. M., Schafer, J. L., & Kam, C.-H. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage.

Enders, C. K. (2001). The impact of nonnormality on full information maximum likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*, 352–370.

Enders, C. K. (2005). Estimation by maximum likelihood. In: B. Everitt & D. C. Howell (Eds), *Encyclopedia of behavioral statistics* (pp. 1164–1170). West Sussex, UK: Wiley.

Enders, C. K. (2006). Analyzing structural equation models with missing data. In: G. R. Hancock & R. O. Mueller (Eds), *A second course in structural equation modeling* (pp. 313–342). Greenwich, CT: Information Age.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430–457.

Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 80–100.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In: K. J. Bryant & M. Windel (Eds), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325–366). Washington, DC: American Psychological Association.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In: R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Kovacs, M. (1992). *Children's depression inventory*. North Tonawanda, NY: Multi-Health Systems.

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiple-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, *86*, 1065–1073.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Montague, M., Enders, C. K., & Castro, M. (2005). Academic and behavioral outcomes for students at risk for emotional and behavioral disorders. *Behavioral Disorders*, *31*, 87–96.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide [Computer software and manual]*. Los Angeles: Muthén & Muthén.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525–556.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, *25*, 99–117.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.

Schafer, J. L. (1999). *NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]*. University Park, PA: Department of Statistics, The Pennsylvania State University.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571.

Shriner, J. G., & Wehby, J. H. (2004). Accountability and assessment for students with emotional and behavioral disorders. In: R. B. Rutherford, M. M. Quinn & S. R. Mathur (Eds), *Handbook of research in emotional and behavioral disorders* (pp. 216–234). New York: Guilford Press.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford.

Wagner, M., Blackorby, J., Cameto, R., & Newman, L. (1993). *What makes a difference? Influences in school outcomes of youth with disabilities*. Menlo Park, CA: SRI International.

Walker, H. M., & Severson, H. H. (1992). *Systematic screening for behavior disorders*. Longmont, CO: Sopris West.

Wilkinson, L. & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In: T. D. Little, K. U. Schnabel & J. Baumert (Eds), *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples* (pp. 219–240). Mahwah, NJ: Erlbaum.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In: M. Becker & M. Sobel (Eds), *Sociological methodology 2000* (pp. 165–200). Malden, MA: Blackwell.

APPENDIX A. M*PLUS* SYNTAX FOR A GROWTH CURVE ANALYSIS WITH AUXILIARY VARIABLES

```
data:
file = ‘c:\cdi.dat’;
variable:
names are
   subcode
   atrisk male speced
   daysunex dayssusp
   basc fcat
   age1 − age7
   cdi1 − cdi7;
usevariables are
   male − cdi7;
tscores = age1 − age7;
missing are all (−99);
define:
male = male * 10;
speced = speced * 15;
fcat = fcat * 5;
basc = basc * 5;
daysunex = daysunex/2;
analysis:
type = random missing;
estimator = mlr;
model:
! BASIC GROWTH MODEL SPECIFICATION;
i s|cdi1 − cdi7 at age1 − age7;
[i s];
i s;
i with s;
cdi1 – cdi7 (1);
! AUXILIARY VARIABLE CORRELATIONS;
cdi1 – cdi7 with male – fcat;
male – fcat with male – fcat;
```

APPENDIX B. SAS MULTIPLE IMPUTATION SYNTAX

```
/* READ RAW DATA */
data speced ;
infile ‘c:\data\cdifinal.dat’;
input
   subcode
   male speced
   daysunex dayssusp
   basc fcat
   time1 − time7
   cdi1 − cdi7;
/* CHANGE MISSING VALUE CODE FROM −99 TO . */
array x[21] subcode – cdi7;
do i = 1 to 21;
if x[i] = −99 then x[i] = .;
end;
drop i;
run;
/* CREATE M = 10 IMPUTED DATA SETS
NIMPUTE SPECIFIES THE NUMBER OF IMPUTED DATA SETS
THE MINIMUM AND MAXIMUM OPTION PROVIDES RANGES
FOR IMPUTED VALUES
ROUND = 1 ROUNDS IMPUTED VALUES TO NEAREST INTEGER
NBITER SPECIFIES THE NUMBER OF BURN IN ITERATIONS
NITER SPECIFIES THE NUMBER OF BETWEEN-IMPUTATION
CYCLES */
proc mi data = speced out = cdimi seed = 56789 nimpute = 10
   minimum = 0 0 0 0 . . . . . . . . . 34 34 34 34 34 34 34
   round = 1 1 1 1 . . . . . . . . . 1 1 1 1 1 1 1;
var male speced daysunex dayssusp basc fcat age1 − age7 cdi1 − cdi7;
mcmc nbiter = 1000 niter = 500;
run;
```

APPENDIX C. M*PLUS* SYNTAX FOR A GROWTH CURVE USING MULTIPLY IMPUTED DATA

```
data:
! FILE AND TYPE ARE USED TO SPECIFY MULTIPLY IMPUTED
DATA;
file = ‘c:\cdifiles.txt’;
type = imputation;
variable:
names are
   subcode
   age1 – age7
   cdi1 – cdi7;
usevariables are
   age1 – age7
   cdi1 – cdi7;
tscores = age1 – age7;
analysis:
type = random;
estimator = mlr;
model:
i s|cdi1 – cdi7 at age1 – age7;
[i s];
i s;
i with s;
cdi1 – cdi7 (1);
```

This page is left intentionally blank

SEEING THE FOREST AND THE TREES: A MORE RIGOROUS APPROACH TO MEASUREMENT AND VALIDITY IN BEHAVIORAL DISORDERS INTERVENTION RESEARCH☆

Maureen A. Conroy and Janine P. Stichter

ABSTRACT

With the national emphasis on the use of evidence-based practices in educational settings, intervention research within the field of special education is being scrutinized. No Child Left Behind (NCLB) has defined evidence-based practices primarily by research that is based on quantitative, experimental designs (i.e., RCT). Although the use of appropriate

☆Development of this paper was partially supported with funding from the U.S. Department of Education, Office of Special Education and Rehabilitative Services (No. H32D990024) and U.S. Department of Education, Office of Special Education Programs, Center on Evidence-based Practices: Young Children with Challenging Behavior (No. H324Z010001). The opinions expressed by the authors are not necessarily reflective of the position of or endorsed by the U.S. Department of Education.

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 131–155

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19006-0

experimental designs has an important place in educational research, defining evidence-based practices based on research design alone is limiting. One critical aspect of research that has not received much attention is the importance of rigorous and precise measurement and systematic replication of research findings. The purpose of this chapter is to review issues surrounding measurement and its effect on validity in intervention research in the field of behavioral disorders. Specifically, we discuss how more rigorous measurement can positively influence the internal, external, construct, and social validity of research findings. A review of current trends in behavioral disorders intervention research is discussed as well as implications for future research.

This chapter is about rigor and measurement within intervention research in the field of behavioral disorders (BD). Specifically, we discuss how the rigor of our measures, not just the design we choose influences the integrity and validity of our research findings. As researchers, we are all *aware* of this basic principle; yet, we observe that intervention research in the field of BD strays at times from strict adherence to this principle. Therefore, the purpose of our chapter is to review issues surrounding measurement and issues related to validity in BD intervention research in an effort to evoke thought and discussion on the issues; thus, influencing future research outcomes. First, we discuss national conceptual and methodological shifts in intervention research and how these shifts have influenced BD research. Second, we discuss the state of BD intervention research with an emphasis on the evolution of measurement and design from a historical perspective. Next, the role of measurement in developing scientifically based practices and the resulting influence of measurement on the validity of our findings are illustrated. Finally, we end with a brief discussion on implications for future research.

CURRENT NATIONAL TRENDS

Over the past few years, national policy has heightened the attention paid to issues related to educational research methodology and design. For example, since 2002, with the enactment of the No Child Left Behind Act (NCLB) (U.S. Department of Education, 2002), educators across the nation have increased their orientation toward the use of *effective, research-based practices*, also referred to as "*evidence-based practices*" or "*scientifically based practices*." NCLB emphasizes the use of educational practices based on quantitative, experimental designs that better establish causal relationships between the variables (U.S. Department of Education, 2002). Hence, increasing attention on identifying educational practices supported by research and "closing the research to practice gap" are a national focus (U.S. Office of Special Education and Rehabilitative Services, 2003). Similar to other educational researchers, many researchers in the field of BD have always been concerned with the dissemination of educational practices based on research; however, with the enactment of NCLB, references to "evidence-based," "scientifically based," and "empirically based" practices substantially increased in the literature and at professional meetings.

With the national attention toward identifying "evidence-based practices," the Institute for Educational Sciences (IES), which developed standards for evaluating the scientific support for current educational practices, was created. IES specifically outlined indicators through the *What Works Clearinghouse* for determining the soundness of educational practices based primarily on research design (Mague, 2004). These indicators clearly present educational research conducted using randomized clinical trials (RCT) with a relatively large number of participants and replications as the preferred methodology (i.e., "gold standard") for identifying an evidence-based practice. As described by Sasso (2005), three standards of research have evolved from NCLB: (1) Gold Standard (i.e., RCT – experimental, randomized), (2) Silver Standard (i.e., quasi-experimental), and (3) Bronze Standard (i.e., supplemental research including correlational, qualitative, and single subject design). An example, of how these standards are implemented can be found in the following illustration. In August 2004, the *What Works Clearinghouse* (2004, August) identified peer assisted learning (P.A.L.S.) as an "evidence-based practice" (www.w-w-c.org). Yet, of the 191 P.A.L.S. studies reviewed, 176 were found to have the scientific rigor to be included; where as, 15 of the studies were found to lack scientific evidence of a high enough standard to be included in the review. Unfortunately, the majority of the studies discarded were single subject or small group studies conducted by the founders and initial researchers in this area[1]. As seen by this illustration, the definition of evidence-based practices and the research to support these practices appears to have either been narrowly defined or remains undefined by IES (Simpson, 2005).

Although the use of RCT certainly has an important place in educational research, defining evidence-based practices based primarily on this design discounts the tremendous amount of educational research conducted using other methodologies, such as quasi-experimental, correlational, qualitative, and single subject-design research. As discussed by Odom (2004), RCT is not an appropriate design for answering all types of educational research questions. Different stages of educational research require different types of designs.

Recently, Odom and colleagues (in press) and others (Levin, O'Donnell, & Kratochwill, 2003), emphasized the need for varied research design and methodology and suggested the following stages of educational research: (1) initial hypothesis and exploration; (2) controlled experiments and demonstrations; (3) randomized clinical trials, and (4) identification of variables adopted for practice, emphasizing the need for varied research design and methodology. Concurrently, a series of papers were published in *Exceptional Children* (see Winter, 2005 issue) outlining quality indicators for defining evidence-based practices and the evaluation of research employing different design methodologies. As a result, increased and ongoing professional discussion has ensued regarding the quality of the methodology and design we use in our educational research. As suggested by Dunst, Trivette, and Cutspec (2002), evidence-based practices can encompass varied research methodologies, establishing relationships between different variables in an objective and credible manner. The issue at hand remains how to operationally and functionally define "evidence-based practices" across those methodologies.

As researchers, we have to ask the question, "Is it important to align our research with national policy?" As many have discussed policy often exceeds our knowledge base and can drive research (Nelson, Roberts, Mathur, & Rutherford, 1999; Sasso, Conroy, Stichter, & Fox, 2001). Although aligning our research with policy can be problematic (especially when the policy is limiting as is the case with the gold standard created by NCLB), as researchers and advocates for the children and youth with behavior disorders, we have a responsibility to provide the empirical evidence that helps determine which policies and practices are effective and which ones are not. We believe this can only be done by continually critically examining the status of our current research base, exploring methodologies that support a more symbiotic relationship between trends in behavioral intervention research and national policies, and employing evidence-based practices that are effective and efficient in serving the population of children and youth with behavioral disorders.

STATE OF RESEARCH IN THE BD FIELD: COMPATIBILITY OR CONFLICT WITH NATIONAL POLICY?

Most behavioral intervention research studies are historically grounded in principles of applied behavior analysis (ABA). In a seminal paper, Baer, Wolf, and Risley (1968) set the stage for the use of single subject design (SSD) as a methodology for examining functional relationships between dependent and independent variables. They emphasized (a) precise measurement of observable behaviors, (b) comprehensive descriptions of behavioral interventions, (c) socially valid research, and (d) generality. This original paper was followed 20 years later by another paper with an increased emphasis on designing single subject studies to address more complex questions, increasing the generality and social validity of findings, and addressing system-wide global interventions (Baer, Wolf, & Risley, 1987). Although written about the use of SSD methodology to study complex interactions between behaviors and the manipulation of environmental factors as an intervention, these two papers address important conceptual foundations for all types of research methodology – (1) precise measurement of dependent measures, (2) a comprehensive description and measurement of independent variables, and (3) conducting socially valid research that has generality across individuals, settings, and time.

Since Baer et al. (1987), the field of ABA has witnessed a dramatic expansion. For example, advances in technology have provided a framework for developing complex data collection systems, allowing us to examine a number of behaviors and environmental factors simultaneously and perhaps more precisely than ever before (e.g., see Tapp & Wehby, 2000**)**. In addition, behavior analysts have employed the use of multiple measures (e.g., indirect and direct) to investigate variables that go beyond the three-term behavioral interaction, such as private events and setting events. Finally, there has been an increase in the use of complex single subject designs (e.g., use of analogue probe methodology, multi-element designs, and combination designs). Although considerable methodological advances have occurred in the field of ABA, these advances have not always translated into intervention research in the field of behavior disorders.

CURRENT PERSPECTIVES OF BD INTERVENTION RESEARCH

Today, two parallel approaches to behavioral intervention research that stem from ABA are present. One approach emphasizes research with individual children or a small number of participants. This approach uses SSD methodology and stresses precise measurement of the dependent variables and visual analysis to determine functional relationships between the dependent measures and independent variables. Although external validity through replication and social validity are critical aspects of this research, due to the small number of participants typically studied, the number of replications required to deem these practices as evidenced-based in applied settings with natural change agents such as teachers has been limited. Examples of this line of research include areas such as functional analysis (e.g., Carr, Newsom, & Binkhoff, 1980; Sasso et al., 1992; Shirley, Iwata, & Kahng, 1999; Stichter, Hudson, & Sasso, 2005), functional communication training (FCT) (e.g., Carr & Durand, 1985; Jolivette, Stichter, & Houchins, 2005), and antecedent-based interventions (e.g., Peck, Sasso, & Jolivette, 1997; Weeks & Gaylord-Ross, 1981; Stichter, Lewis, Johnson, & Trussell, 2004). A second approach to research examines the use of ABA interventions from more of a systems change perspective. Although exceptions exist, these research studies are more likely to examine large groups of participants or use the classroom or school as the unit of analysis, use indirect measures (e.g., discipline referrals, behavioral rating scales), and employ group design methodology (e.g., correlational, or quasi-experimental designs) to examine the relationships between the dependent measures and independent variables. Although establishing a strong relationship between the dependent measures and independent variables is an important part of this research, due to the use of indirect measures and quasi-experimental designs, the exact nature of this relationship is often less clear. One strength of these studies is the ability to demonstrate the potential impact of interventions across a large number of participants in applied settings; thus, emphasizing the social validity and potential for multiple replications of the intervention. Examples of this line of research include positive behavior supports (PBS) (Horner, Sugai, Todd, & Lewis-Palmer, 2005; Safran & Oswald, 2003), social skills groups (Lewis, Powers, Newcomer, Johnson, & Bradley, 2003; Gresham, Sugai, & Horner, 2001), and anti-bullying intervention programs (Elliott, 2002; Olweus, 1993; Walker, Ramsey, & Gresham, 2004). Although both approaches address important research questions, often times, the questions they address are different and ergo, emphasize various aspects, utilization of designs, and stages of research.

ALIGNMENT WITH NATIONAL POLICY

Current national educational research policies as described by the Institute of Educational Sciences (IES) allude to different stages of research; viewing these stages predominately from a developmental model. At this time, given the current state of BD intervention research, it is unclear to what degree either of the two approaches described above can flourish under such a model. For example, in the May 6, 2005 Federal Register, the IES published a notice inviting applications for grants to support educational research. In this request for proposals (RFP), IES clearly outlined a research template with four developmental stages of research, identified as separate "goals" and a fifth stage (goal) oriented toward assessment instruments, which applicants could apply for within their specific content areas (e.g., serious BD). Goal 1 was primarily oriented toward analyzing existing data sets in order to assess current practices, with the anticipation that the analysis will then set the stage for other goals (e.g., Goal 3, applications for next steps in research and practice). Goal 2 was designed for the development of educational programs and practices that may show promise, with the expectation that they can be later developed as Goal 3 applications. Under Goal 2, both SSD and RCT designs were deemed appropriate to the degree that they were relevant to the research questions. Goal 3 was designed for replication and efficacy trials of educational programs and practices. In this goal, RCT designs were indicated as the preferred design; however, SSD designs were considered a possible alternative, requiring justification. Goal 4 was designed to go to scale with the educational programs and practices and demonstrate efficacy and effectiveness in community settings and for the typical persons who would be influenced. Randomized field trials and some forms of quasi-experimental designs were only permitted due to the scope of the desired outcomes.

Although a developmental progression of research as outlined by IES is logical for some areas of research, this model implies a type of "one size fits all" approach to answering various hypotheses at various stages of research. Given the individuality of many behavioral intervention strategies, such as FCT or functional analysis, conducting large N-studies that use random clinical trials may be less appropriate. For example, researchers may be able to demonstrate the effectiveness of the use of an individualized intervention procedure, such as FA or FCT, on a large group of students; however, the outcome measures for these interventions cannot by design be standardized across a large group of students, and therefore, outcomes for individual students within that group are likely to differ, preventing researchers' ability to make generalized conclusions about specific outcomes for the entire group. To illustrate, consider the application of functional analysis. Functional analysis has a long history in the field of ABA with ample evidence supporting its use to address problem behaviors. Yet, functional analysis produces highly individualized results – that is, for each individual a different function may be identified. In addition, functional analysis researchers have found that for a number of individuals functions are not identified through the use of this strategy (for a review, see Asmus et al., 2004). Because FA may not always produce consistent results across all individuals does not mean that this strategy is not useful or effective for many individuals. In fact, often times, the results of an inconclusive FA can help guide an interventionist to examine other contextual variables influencing the behavior, establishing operations or antecedents and incorporate those variables into an effective intervention (e.g., see Carr, Yarbrough, & Langdon, 1997). Therefore, functional analysis may indeed be an evidenced-based intervention, but one that is intended to be employed with individuals, rather than large groups. For this reason it has most commonly been identified as a tertiary intervention best implemented on an individual basis for those who do not effectively respond to large group interventions. As a result, conducting research in a developmental progression as suggested by IES may not be appropriate for an individualized intervention strategy, such as functional analysis. Rather, continued use of appropriate and rigorous measures matched with the optimal research design that answers the research question, replicated across multiple individuals and research sites would seem be a more logical progression.

The parallel conceptual shift in behavioral intervention research discussed earlier emphasizing evidence-based practices that are able to incorporate larger number of participants using a large-scale systems approach poses equal research complexities when considering the stages of research as described in current IES policies. To illustrate, school-wide positive behavior support (PBS) is an example of this shift in the field of BD. Although, in essence, PBS is not considered as an evidence-based practice in and of itself, PBS has been described as an applied science (Carr et al., 2002) based on principles of research (Scott, Liaupsin, Nelson, & McIntyre, 2005). A significant number of research studies evaluating the effectiveness of behavioral intervention practices within a PBS framework have been conducted (for a review, see Carr et al., 2002). However, increasingly in this area of research within the BD field, the "participant" is often not the individual child, but is more likely to be the classroom or the school (for a discussion, see Kern & Manz, 2004). For example, much of the school-wide PBS research would be most closely aligned with IES Goals 3 and 4, described as an intervention process implemented on a wide-spread basis in applied settings. The optimal outcome of school-wide PBS is to provide staff in schools an intervention framework or process that contributes to the prevention of school-wide problems behaviors, while promoting prosocial and academic outcomes. Additionally, the focus is on decreasing problem behaviors in groups of students across typical school-wide contexts, which theoretically may reduce the need for more individual interventions with students. However, establishing a strong functional relationship between behavior change across groups of students and school-wide PBS strategies implemented has been problematic in the literature (Kern & Manz, 2004). For example, global measures, such as office referrals, school-wide surveys on the effectiveness of PBS, and school-wide academic assessments, provides limited understanding of the functional relationship between specific PBS strategies and associated outcomes for students' problem behaviors. Global measures associated with larger scale research, such as school-wide PBS, typically derive their strength from, either previous smaller scale research that provide repeated demonstrations of a functional relationship between specific strategies and behavior change, such as SSD, or randomized field trials across a large group of students using a multi-component treatment package. In contrast, initial school-wide PBS research utilizing these measures has relied primarily on descriptive data; thus, creating threats to the internal validity of the current research findings in this area. Additionally, the components of PBS as a multi-component package has not been explicitly described making it difficult to replicate the effects of school-wide PBS strategies across investigators and research sites. Since PBS is comprised of many isolated strategies, technically, IES Goals 1 and 2 may be most applicable for investigating and analyzing the functional relationship of isolated school-wide PBS intervention strategies prior to fully implementing PBS as a school-wide process. Given that the evidence supporting the use of school-wide PBS is based primarily on quasi-experimental design methodology, in its present form, school-wide PBS research does not appear to be aligned with the national emphasis on RCT and would most likely also not be identified by the *What Works Clearinghouse* as an evidence-based practice. This is not to say that PBS, functional behavioral assessment, inclusion, and other initiatives may not have produced beneficial outcomes for teachers, schools, children, and families, but rather, these initiatives lack the precision in measurement and design to be aligned with current policy defining evidence-based practices. Thus, making it difficult to know, which components of each practice works in various settings and with whom.

The two approaches to BD intervention research highlighted separately address a continuum of research methodologies and current initiatives. The first approach (i.e., individual child or small N interventions) typically uses methodology that emphasizes establishing a functional relationship between the dependent and independent variables, while examining external and social validity through systematic replication; where as, the second approach (i.e., large group or system interventions) establishes a statistical relationship between the dependent measures and the intervention, but accentuates the social validity of the practices. Each approach to behavioral intervention research has strengths, but neither approach is sufficiently comprehensive to align with current national policies.

Idiosyncratic Measurement and Design Issues in BD Intervention Research

Measurement of dependent variables in BD intervention research is different than research in other disability areas (e.g., research examining reading abilities) (see Forness, 2005, for a discussion). Presently, in BD research we lack standardized measures that are often sensitive to the behavioral construct being measured (Stichter & Conroy, 2004). The standardized measures available, and often used in research studies to measure problem and adaptive behaviors (e.g., *Child Behavior Checklist* (*CBCL*); Achenbach, 1991), were developed primarily for identification purposes and not necessarily designed to provide measures sufficiently sensitive for purposes of intervention research. The idiosyncrasies of measuring problem and adaptive behaviors stem primarily from the topographical nature of the dependent variables typically measured in BD intervention research. Problem behaviors, such as disruption, aggression, noncompliance as well as adaptive behaviors, such as compliance, on-task behavior, social competence, and engagement are not as easily quantified as an academic skill, such as reading and math (Forness, 2005). In addition, problem and adaptive behaviors are influenced by context, specifically the immediate behavioral events that surround that behavior. Therefore, measurement of problem and adaptive behaviors can be difficult and often need to be individualized to account for the contexts in which they occur. Consider a dependent measure such as "compliance." As a behavior, "compliance" can be quantified (e.g., target child responds to or completes an adult request within 5 s). Yet, unlike an academic skill, which has not been taught, this operational definition may not reflect the reason why the student does not comply. Rather, it simply represents the topography by which the desired outcome might be measured. When students read fluently, the assumption is that they have acquired a set of skills that are considered desirable for future success across other academic as well as functional life skills. Whereas, in reading most students go through common stages to become literate (e.g., awareness, acquisition, fluency); however, for students with BD the prerequisite skills for skills such as compliance are far from standardized. Some students may need to learn an academic skill in order to comply with a teacher's request; other students may need to learn contingencies for engaging in compliant behavior as opposed to noncompliance. Moreover, a target child's compliance to a request may be highly dependent on the individual who is requesting, the specific task or skill requested, the target child's ability level to comply to the request, the setting in which the request is made, potential setting events that occurred prior to the request, and the availability of reinforcers following the request. Measuring or controlling for all of these environmental considerations is difficult to say the least. Yet, if we fail to consider the impact of these environmental factors on the behavior or we use a teacher rating measure to assess compliance, we may not accurately measure the behavioral construct of "compliance." For example, if we measured compliance using a Likert-type rating scale, the respondent may rate compliance differently over time depending on the day the scale was completed, the activities that occurred that day, the level of the target child's compliance that day, opportunities for the child to comply and so forth. Of course, there are ways to help control for respondent bias (e.g., inter- and intra-rater reliability); however, we rarely see this level of measurement in studies that primarily use indirect measures.

As a result of these measurement complexities, many BD intervention researchers choose to use precise measures along with SSD methodology. Although precise measurement is an overall strength of SSD methodology, SSD design clearly has limitations – resources, time and, in some cases, practicality often limits sufficient and necessary systematic replication to ensure external validity of the findings. In other studies, where between-group design has been employed, researchers have a tendency to rely on the same, overused indirect measures, which may not provide the detailed measurement to determine a functional relationship and help guide interventions – that is, knowing the effects of a specific intervention strategy on a particular behavior. BD intervention research is now at an intersection with current national policy and the current literature base, where increased emphasis on the development of appropriately sensitive dependent measures as well as replication across larger groups of participants, settings, and strategies must occur in order to execute meaningful and effective results.

In sum, precise measurement is needed to accurately represent the construct of many problem and adaptive behaviors investigated in BD intervention research, such as compliance, aggression, disruption, engagement, and social competence. However, when precise measurement of these behaviors occurs using SSD methods, the outcomes can be limiting. In addition, optimal measurement can be problematic when designing large between group research studies due to the lack of available precise, standardized behavioral measures that have construct validity (Forness, 2005). As suggested by a number of researchers, a multi-source measurement paradigm that includes both precise measures of dependent variables as well as global measures of the construct of emotional and behavior disorders may be the most appropriate solution to this issue (for a discussion, see Conroy, Hendrickson, & Hester, 2004). Regardless of methodology and design used, there is a need to conduct studies that emphasize both micro (i.e., studies with high levels of construct and internal validity) and macro (i.e., studies with high levels of social and external validity) analyses to improve problem behaviors in children and youth with behavioral disorders. Researchers who focus primarily on measuring precise behaviors may only be evaluating the ‘‘trees’’ and not the ‘‘forest.’’ Whereas, researchers who measure variables that only represent the ‘‘forest’’ may be missing the effects of the intervention on the ‘‘trees.’’ In the next section of this chapter, we discuss how using rigorous measurement of both the dependent and independent variables (regardless of the design employed) can improve the quality of all BD intervention research, increasing the construct, internal, external, and social validity of practices used.

THE RELATIONSHIPS BETWEEN MEASUREMENT, VALIDITY, AND EVIDENCE-BASED PRACTICES

Science is the process that describes, predicts, or controls a phenomenon of interest (Sidman, 1960). Science is not determined by the number of participants or the type of design employed, but rather by actual replication of the findings (Herschbach, 1996). Rather, measurement and control are the key features of science (Stanovich, 2004). As discussed by Sasso (2005), science is a process of systematic empiricism that includes reliability, replication, and validity. The scientific process does not *prove* a phenomenon, but rather *describes or predicts* the outcomes and eliminates error (Sasso, 2005).

We contend that *precise measurement* along with systematic replication should be one of the ‘‘gold standards’’ for conducting research and defining evidence-based practices in behavioral intervention research, rather than reliance on a particular experimental design or model. If researchers employ strong measurement systems across both dependent and independent variables and account for rival hypotheses, stronger validity will ensue regardless of the designs used. In this section, we discuss the impact of measurement on all types of validity: construct, internal, external, and social.

CONSTRUCT VALIDITY

Construct validity is most often discussed in relation to standardized psychological testing defined as a measure of an attribute or quality that accounts for variance in a test instrument (Cronbach & Meehl, 1955); thus, indicating that the "test" accurately represent the construct or attribute of interest. Although this traditional definition of construct validity is commonly used in the literature, construct validity is applicable to a broader set of measures than just standardized psychological tests. As suggested by Bechtoldt (1951), construct validity "involves the acceptance of a set of operations as an adequate definition of whatever is to be measured" (p. 1245). Therefore, we suggest that behavioral intervention researchers consider the construct validity of the dependent and independent variables under investigation – that is, do the measures used accurately represent the constructs they purport to measure? Due to the lack of standardized behavioral measures available in the field of BD, at times, researchers may choose to use dependent measures in their research that lack the construct validity for which they are intended. Consider a research study that investigates the effects of a specific behavioral intervention on children's problem behavior. To measure the effects of the intervention on problem behavior the investigators have chosen to use the *Systematic Screening for Behavior Disorders* (*SSBD*; Walker & Severson, 1992). Although the SSBD is a valid measure for use as a screening tool to identify children and youth who are at elevated risk for emotional or behavioral disorders, this instrument has not been validated for use as a standardized measure of occurrence of "problem behavior" or changes related to intervention. Because the SSBD is not designed for this purpose, its use in this investigation as a dependent measure is questionable.

In addition to the lack of standardized measures available for behavioral intervention researchers, an additional concern is the lack of a standard definition of the construct of *behavior disorders and problem behavior*. Given that behavior is a context specific and fluid construct, one researcher may design an intervention to address *problem behavior representative of a behavioral disorder* defined in one particular manner; whereas, a second researcher may be using a different definition. This concern is also present as we define not only our dependent variables, but also our independent variables in behavioral research. For instance, when we examine the effectiveness of an independent variable, such as FBA, on the assessment and treatment of problem behaviors, we need to make certain that we not only operationally define and obtain procedural integrity for the independent variable (i.e., the components of FBA), but also define and measure the construct of *problem behavior* with consistency, precision, and accuracy. We need to verify that the measures we use to evaluate changes in the dependent variable (e.g., problem behaviors) are consistently well-defined across studies with precision and sensitivity to provide an accurate and valid measure of the construct. For example, when measuring problem behaviors, how researchers define these behaviors can significantly impact the generality of findings. Consider the impact of two different types of definitions and measures of the construct of *problem behaviors* as illustrated in two hypothetical studies. In one study, problem behaviors are defined as any behavior that leads to an office discipline referral and measurement of problem behaviors is the total number of office referrals reported in a school before and after treatment. At the end of this study, the researchers indicate their specific intervention strategy reduced ‘‘problem behavior’’ as measured by a decrease in office referrals. In another study, problem behaviors are very specifically identified as defiance/noncompliance and disruption and operational definitions are developed. The rate of these problem behaviors is systematically measured for each individual participant in the school before and after treatment. At the end of this study, the researchers conclude that their intervention reduced problem behaviors as well. Although both definitions may be valid in addressing the particular research questions of interest in each study, their definition and measurement of *problem behavior* significantly differed across both studies, which influences the generality of their findings. Clearly, how problem behaviors are defined and measured provides different levels of construct validity that influence research findings. Do the number of office referrals accurately represent the construct of ‘‘problem behaviors.’’ If so, which problem behaviors are represented by this measure? Are office referrals a reliable and valid measure of problem behavior? If so, can the accuracy of measuring the construct of problem behaviors using office referrals be enhanced to increase the construct validity of this measure? The use of office referrals to measure the effects of an intervention is just one example to illustrate our point regarding the construct validity of BD intervention research; however, many other examples exist in the literature (e.g., inclusion). Our point is, that as researchers select measures to represent dependent and independent variables, these measures should accurately measure the construct they purport to measure.

INTERNAL VALIDITY

Internal validity is the ability to reliably predict the effects of the independent variable on the dependent variable. Therefore, selecting a measurement system that not only accurately represents the construct, but one that is sensitive enough to reflect changes in the dependent variable(s) is imperative for sound research. For instance, in order to determine if the chosen behavioral intervention reliably predicts any changes in the dependent variable, the measurement system needs to be precise enough to truly evaluate these changes. Once again, consider the illustration of the use of office referrals as a measure of problem behavior. Do decreases in the number of office referrals following the implementation of the independent variable over a specified period of time provide accurate evidence of actual decreases in problem behaviors? Is there a way to develop a more sensitive measurement system with office referrals as the unit of measure for individual children or for specific infractions? Also, consider the following illustration of the use of a direct measurement that may lack sensitivity – interval recording. Much of behavioral intervention research uses different types of interval recording procedures to measure changes in behavior. Whether the measurement system uses partial interval, whole interval, or momentary time sampling, interval recording procedures can lack sensitivity, particularly if one is measuring high- or low-rate behaviors or behaviors that do not lend themselves to this type of measurement system (e.g., compliance) (Repp, Roberts, Slack, Rapp, & Berkley, 1976). Simultaneously, precise and consistent measures regarding the integrity of the implementation of the independent variable better supports that the intervention of interest is responsible for changes in the dependent variable. This is particularly relevant in applied research within schools where multiple intervention initiatives and contexts are frequently co-occurring. Simply put, the more precise and reliable the measure, the more accurate the research findings, thus minimizing the chance of type 1 or type 2 errors and increasing the internal validity of the research.

EXTERNAL VALIDITY

External validity is the ability to extend research findings to other individuals with different or similar characteristics, behaviors, settings, and across time. This is accomplished through sufficient replication as opposed through employing a particular design or including a large number of participants in a single research study. In addition to internal validity, external validity is a vital part of the scientific process. Without external validity, the findings of research are exceedingly restricted. Although a single experimental verification of a particular intervention on an individual participant or small group of participants has value, if this same behavioral intervention is either not further investigated, not effective in promoting change, or lacks durability across other participants' behaviors, settings, or time, the external validity of the practice is compromised. Of course, measurement is a fundamental component of replication and is the essence of generality and external validity. All too often, SSD research fails to measure and design systematic replications of a practice. Horner et al. (2004) suggested that replication using SSD should occur across five SSD studies with a total of 20 participants across research teams and sites to be considered an evidence-based practice. Although a formal evaluation of examining practices based on SSD studies has not been conducted to date, a preliminary analysis on academic intervention research for students with BD suggests that few practices meet these criteria (Hudson, 2005) (though exceptions do exist). Similarly, group design research studies all too often fail to measure and conduct systematic replications as well.

It seems once a functional relationship has been established between the practice and the behavior change for various reasons, systematically evaluating the external validity of the practice seems to be given less attention. If indeed this is the case, as behavioral researchers we need to increase our systematic replications of research findings. The first step to increasing the external validity of our research is to assure that we are measuring practices across representative samples of participants, behaviors, settings, and time. To illustrate this point, consider the existing functional analysis research base. The effectiveness of functional analysis as an assessment tool for identifying the functions of problem behaviors has been validated across a number of individuals with severe disabilities within clinical settings (Iwata, Dorsey, Slifer, Bauman, & Richman, 1982/1994; Northrup et al., 1991). Even though some replication studies have been conducted, this same practice has far less data to support its use with children and youth with BD in applied settings with natural change agents (for a discussion, see Sasso et al., 2001). Therefore, although functional analysis research probably technically meets the criteria outlined by Horner and colleagues, the external validity of the practice for individuals with behavior disorders in applied settings is far less clear. Plainly, increasing measurement of practices and replication of findings across representative samples of participants, settings, behaviors, and time should be addressed to improve the external validity of behavioral intervention research.

SOCIAL VALIDITY

Conducting research on behavioral intervention strategies that can be implemented in applied settings and that improve the quality of life for individuals with behavior disorders is the essence of developing socially valid practices that bridge the research to practice gap. Again, measurement is a key factor in developing the social validity of behavioral intervention research findings. For example, using social validity measures that are meaningful and produce reliable information on the validity of the practice under investigation for the individuals with behavior disorders and their peers, parents, and teachers is imperative for developing valuable intervention strategies. However, most often, social validity measures are comprised of a Likert-type scale and are often added onto an existing study and under emphasized. Although these measures can produce important information, in their current forms, many are not particularly meaningful in understanding and evaluating the social validity of the practice. To obtain precise measures of social validity, researchers will need to consider using both direct and indirect measures. For instance, to determine if the individual's behavior improved in a socially meaningful way following intervention, researchers need to implement direct measures that evaluate the social implications of the individual's behavior before and after treatment. In addition, direct and indirect social validity measures can be obtained by persons in the individual's community, school, and family. Returning to the seminal paper by Baer et al. (1987) mentioned previously, research needs to be replicated to have generality across individuals, settings, and time. In addition, precise measurement of social validity should be incorporated as replication of research occurs, assuring that the practices under investigation are socially valid across various contexts. Therefore it is crucial, in the identification of evidence-based practices, to design measures that can directly target these questions, such as those that assess improved outcomes in ways that are considered meaningful for the relevant contexts.

In summary, without precise measurement and strong construct, internal, external, and social validity, the ability to develop evidence-based practices that can be used to help children and youth with BD is limited. In addition, all types of validity are interconnected and sacrificing one type of validity influences other types of validity. For example, broad indirect measures of constructs may decrease the construct and internal validity of findings. In the same respect, precise measurement of behaviors in isolated, controlled conditions may decrease the social and external validity of findings. Furthermore, defining external validity by replication alone does not necessarily produce generality of findings. Although replication is a critical component for increasing the external validity of research, if the internal or construct validity of the research is weak, the number of studies or participants included in investigations does not, in and of itself, lead to external validity.

Finally, we contend that not all types of validity should be treated equally. Recently, four additional types of validity were conceptualized and applied to educational research to assess the value of research as it related to political and public perception (Fabes, Matrin, Hanish, & Updegraff, 2000; Gersten et al., 2005). These four types of validity are: (1) incidence validity (i.e., the degree to which a topic effects large number of individuals), (2) impact validity (i.e., the degree to which the topic is thought to have significant consequences), (3) sympathetic validity (i.e., the ability of the topic to generate emotion), and (4) salience validity (i.e., the degree to which the public is aware of the topic) (Gersten et al., 2005). These types of validity may help to explain how and why some types of research are prioritized and funded and other types of research are not prioritized, irregardless of methodological soundness of the research design. However, they do not provide direction for strengthening the findings of the research or implications for addressing a particular phenomenon of interest, particularly interventions. Actually, these forms of validity do not even reliably determine a problem of concern – that is, they appear to suffer from their own form of construct validity. For example, if you consider the topic of school violence defined as the incidence of school shootings or weapons offenses that occurred toward students in schools and assessed it against all four of these types of validity, you would be assured that school violence is a very pressing problem in today's schools. Yet, actual research data indicates the rate of school shootings and weapon offenses has decreased (NCES, 2004). Media coverage of school violence, however, has increased.

FUTURE RESEARCH DIRECTIONS

Precise measurement and appropriate research designs are the foundations for developing scientifically based practices. As discussed, measurement influences the validity of our findings including construct, internal, external, and social validity. Rather than focusing on a preferred research design, our focus needs to move to developing precise measurement systems that can accurately reflect our findings. Therefore, measurement systems should document experimental control (i.e., internal validity) by precisely measuring the targeted behaviors and the behavioral intervention. For example, if we say that using strategy XYZ decreases problem behavior, we need to specifically describe, measure, and report findings on the problem behavior we are changing (e.g., disruption, physical aggression, noncompliance). Given our current indirect measures (or lack of standardized measures) and the problems with indirect measures (e.g., bias'), this may be more difficult to do. We recommend using a combination of measures including direct measures on specific behaviors whenever possible. We also need to be cautious when stating findings (e.g., if we report that problem behavior decreases, which does not necessarily mean that appropriate behavior increases unless we have measured the construct of appropriate behavior and have the data to support that outcome).

Measurement systems also need to be congruent with the design accurately documenting the change in the phenomenon of interest (i.e., construct and internal validity). Given the extensive foundation of ABA and our consistent use of SSD, BD intervention research is well beyond the point where our interest is in simply demonstrating that we can manipulate behavior. Simply documenting, for example, that a child's appropriate behavior has increased and an inappropriate behavior decreased during a study is no longer sufficient.

It is essential that the measures are related to the areas of interest. For example, research investigating whether the mainstreamed classroom is the best placement for a student to reach his or her targeted educational and or behavioral goals must include specific measures that capture those specific types of goals (as opposed to only relying on indirect measures, such as "success in the inclusive placement"). Furthermore, we should address if the data are trustworthy, defined as a result of a contextually relevant intervention that was properly documented and implemented and supports a direct link between the change in the target behavior and the intervention. Only with this level of treatment integrity can most intervention research for students with BD have the potential for small and large-scale replication as well as future efficacy analysis.

To accomplish this, we must describe and directly measure both the components and the implementation of the intervention (i.e., treatment integrity) to address all aspects of validity. For example, with research in the area of inclusion, we need to accurately describe the intervention that occurred to understand the relationship between this intervention and the change in the dependent measure. How much time was targeted in the mainstream setting to call it "inclusion?" Was this goal achieved? Were ancillary supports, formal or informal peer supports, and levels of academic modifications provided? Were behavioral supports necessary? All of these questions and more must be operationalized and then measured to evaluate the success of a student's placement in an inclusive setting. Otherwise inclusion will remain a philosophical approach to intervention, void of any prominent place on the continuum of meaningful interventions for students with BD. Similarly, PBS has now expanded into a global term, used by such a diverse population of researchers and practitioners, that it can no longer be used within intervention research as an assumed processes defined by the initial investigators. Specific PBS procedures or guidelines will need to be more forthcoming, defining the practice. Such exacting standards are required of any "science."

By enhancing the science by which we measure, document, and report our intervention findings, we can increase our external validity by facilitating replication of findings across participants, behaviors, times, and settings. Replication is the key to external validity, but we need to systematically replicate by (1) precisely identifying the characteristics of the different participants, (2) operationally defining specific behaviors and examining the hierarchical relationships between behaviors and behavioral classes (e.g., noncompliance may be a precursor to aggression, therefore, a decrease in noncompliance may ultimately decrease aggression), and (3) measure and describe the characteristics of the context in which interventions can be successfully implemented (e.g., time of day, type of settings, teacher characteristics, and so forth). For example, with FA research, we need to accurately describe under "what conditions" this type of research can be implemented in order to understand the limitations of the practice.

As highlighted earlier, all types of validity are important and interconnected. Unfortunately, social validity is often given less attention than other types of validity and only evaluated by a Likert-type scale. This information may be useful, but more attention is needed to the measurement of social validity. We need to directly measure whether any socially valid changes occurred for the individual(s) through more precise and targeted measures (e.g., before and after ratings by naïve observers) that recognize social validity as directly related to external validity as well as large scale analyses of efficacy and efficiency.

In summary, high-quality research that accurately defines and reliably measures the constructs that are purported to be measured is an essential part of developing scientifically based practices that are meaningful for children and youth with behavioral disorders. In addition, researchers need to carefully choose measurement systems and design studies that account for rival hypotheses. Finally, systematic replication of future research needs to occur to increase external validity by conducting investigations across participants and settings and to include stages of research.

CONCLUDING THOUGHTS

Most of the issues discussed in this chapter are not new. These are issues that are discussed in most research methodology courses. Developing measures that have high construct validity and maintaining high internal, external, and social validity is always a balancing act when conducting applied research. The purpose of this chapter is to reflect on the effects of measurement systems on the validity of research findings. As the national emphasis on RCT and group design research continues and a parallel emphasis on socially valid systems research is stressed, we suggest that behavioral intervention researchers be wary of joining the bandwagon or abandoning the principles of our ABA roots, without first examining and developing measurement systems that enhance the construct, internal, external and social validity of their findings. Regardless of design, rigorous measurement and control of all components of the scientific process is needed to develop and disseminate practices that include scientific evidence to support their effectiveness in improving the behaviors of children and youth with behavioral disorders. Without increased rigor in measurement, though we may know what an evidence-based practice is, we will continue to be uninformed as to the optimal conditions for its effectiveness or with whom it will be most effective. The field of BD needs to know "what works" and for whom (Guralnick, 1997). For example, we *know* P.A.L.S. works, yet would some of the studies rejected in the IES review have given us strong indications regarding specific learning characteristics of the individuals for whom P.A.L.S. works? We feel to assume that all fairly complex phenomena (i.e., instructional practices) can automatically or eventually be simplified to a process that can be randomly applied and functionally measured is highly problematic. Perhaps, there is gold standard for those processes by which we create data and another for those that condition how we use it. Through implementation of research studies with high-quality designs that include precise measurement systems, the field of BD research will continue to evolve; hopefully, setting congruent "gold standards" of its own.

NOTES

1. As of September 2005, P.A.L.S was removed from the *What Works Clearinghouse* website.

ACKNOWLEDGMENT

We would like to thank William H. Brown, James J. Fox, and Gary Sasso for their thoughtful comments on a previous draft of this chapter.

REFERENCES

Achenbach, T. M. (1991). *Manual for the child behavior checklist/4–18 and 1991 profile*. Burlington, VT: University of Vermont Department of Psychiatry.

Asmus, J. A., Ringdahl, J. E., Sellers, J. A., Call, N. A., Andelman, M. S., & Wacker, D. P. (2004). Use of a short-term inpatient model to evaluate aberrant behavior: Outcome data summaries from 1996 to 2001. *Journal of Applied Behavior Analysis*, *37*, 283–304.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*, 91–97.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *20*, 313–327.

Bechtoldt, H. P. (1951). Selection. In: S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1237–1267). New York: Wiley.

Carr, E. G., Dunlap, G., Horner, R. H., Koegel, R. L., Turnbull, A. P., Sailor, W., Anderson, J. L., Albin, R. W., Koegel, L. K., & Fox, L. (2002). Positive behavior support: Evolution of an applied science. *Journal of Positive Behavior Interventions*, *4*, 4–16.

Carr, E. G., & Durand, V. M. (1985). Reducing behavior problems through functional communication training. *Journal of Behavioral Education*, *18*, 111–126.

Carr, E. G., Newsom, C. D., & Binkhoff, J. (1980). Escape as a factor in the aggressive behavior of two retarded children. *Journal of Applied Behavior Analysis*, *13*, 101–117.

Carr, E. G., Yarbrough, S. C., & Langdon, N. A. (1997). Effects of idiosyncratic stimulus variables on functional analysis outcomes. *Journal of Applied Behavior Analysis*, *30*, 673–686.

Conroy, M. A., Hendrickson, J. M., & Hester, P. (2004). Prevention and intervention of emotional/behavioral disorders in young children. In: R. Rutherford, S. Mathur & M. Quinn (Eds), *Handbook of research in behavioural disorders* (pp. 199–215). New York: Guilford Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Dunst, C. J., Trivette, C. M., & Cutspec, P. A. (2002). Toward an operational definition of evidence-based practices. *Centerscope*, *1*, 1–10.

Elliott, D. S. (Ed.) (2002). *Blueprints for violence prevention: Bully prevention program*. Boulder, CO: Institute of Behavioral Science, Regents of the University of Colorado.

Exceptional Children Special Issue (2005), *71*, 135–207.

Fabes, R. A., Matrin, C. L., Hanish, L. D., & Updegraff, K. A. (2000). Criteria for evaluating the significance of developmental research in the twenty-first century: Force and counterforce. *Child Development*, *71*, 212–221.

Forness, S. R. (2005). The pursuit of evidence-based practice in special education for children with emotional or behavioral disorders. *Behavioral Disorders*, *30*, 309–328.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Gresham, F. M., Sugai, G., & Horner, R. H. (2001). Interpreting outcomes of social skills training for students with high-incidence disabilities. *Exceptional Children*, *67*(3), 331–344.

Guralnick, M. J. (Ed.) (1997). *The effectiveness of early intervention*. Baltimore: Brooks.

Herschbach, D. (1996). Imaginary gardens with real toads. In: P. Gross, M. Levitt & M. W. Lewis (Eds), *The flight from science and reason*. Baltimore: Johns Hopkins University Press.

Horner, R. H., Sugai, G., Todd, A., & Lewis-Palmer, T. (2005). School-wide positive behavior support: An alternative approach to discipline in schools. In: L. Bambara & L. Kern (Eds), *Individualized supports for students with problems behaviors* (pp. 359–390). New York: Guilford Press.

Hudson, S. (2005). *Identification and teacher perceptions of academic practices: Initial results of quality indicators and a three-tiered classification framework*. Unpublished manuscript.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2004). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–180.

Iwata, B., Dorsey, M., Slifer, K., Bauman, K., & Richman, G. (1982). Toward a functional analysis of self-injury. Analysis and Intervention in Developmental Disabilities, *3*, 138–148, 198.

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, D. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis*, *27*, 197–209.

Jolivette, K., Stichter, J. P., & Houchins, D. E. (2005). *Functional communication training for a student with emotional and behavioral disorders*. Unpublished manuscript.

Kern, L., & Manz, P. (2004). A look at current validity issues of school-wide behavior support. *Behavioral Disorders*, *30*, 47–59.

Levin, J. R., O'Donnell, A. M., & Kratochwill, T. R. (2003). Educational/psychological international research. In: W. Reynolds & G. Miller (Eds), *Handbook of psychology, educational psychology* (Vol. 7, pp. 557–581). Hoboken, NH: Wiley.

Lewis, T. J., Powers, L., Newcomer, L., Johnson, N., & Bradley, L. (2003). Implementing small group and individual behavior support plans in the context of a school-wide system of positive behavior support. Paper presentation at the 29th Annual International Association for Behavior Analysis Convention, San Francisco, CA.

Mague, T. (2004). Determining what works: An interview with Dr. Grover Russ Whitehurst. T.H.E. Journal, 32–37.

National Center for Educational Statistics. (NCES). (2004). Indicators of school crime and safety: 2004. U.S. Department of Education: Bureau of Justice Statistics.

Nelson, J. R., Roberts, M., Mathur, S. R., & Rutherford, R. B. (1999). Has public policy exceeded our knowledge base? A review of the functional behavioral assessment literature. *Behavioral Disorders*, *24*, 169–179.

Northrup, J., Wacker, D., Sasso, G., Steege, M., Cigrand, K., Cook, J., & DeRaad, A. (1991). A brief functional analysis of aggressive and alternative behavior in an out patient clinic setting. *Journal of Applied Behavior Analysis*, *24*, 509–522.

Odom, S. L. (2004). The RCT gold standard: Beware of the Midas touch. *FOCUS on research. Newsletter of the Division for Research*, *17*(1), 1–2.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R., Thompson, B., & Harris, K. (in press). Research in Special Education: Scientific methods and evidence-based practices. Education and Treatment of Children.

Olweus, D. (1993). Bullying at school. Malden, MA. Blackwell Publishing. OSEP Technical Assistance Center on Postivie Behavioral Internvetions and Supports http://www.PBIS.org

Peck, J., Sasso, G. M., & Jolivette, K. (1997). Use of the structural analysis hypothesis testing model to improve social interactions via peer-mediated intervention. *Focus on Autism and Other Developmental Disabilities*, *12*, 219–230.

Repp, A. C., Roberts, D. M., Slack, D. J., Rapp, C. F., & Berkley, M. S. (1976). A comparison of frequency, interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis*, *9*, 501–508.

Safran, S. P., & Oswald, K. (2003). Positive behavior supports: Can schools reshape disciplinary practices. *Exceptional Children*, *69*(3), 361–373.

Sasso, G., Conroy, M. A., Stichter, J., & Fox, J. J. (2001). Slowing down the bandwagon: The misapplication of functional assessment for students with emotional and behavioral disorders. *Behavioral Disorders*, *26*, 282–296.

Sasso, G. (2005). The evidence base in emotional and behaivoral disorders. Presentation at the University of Florida Research Symposium. Gainesville, FL.

Sasso, G. M., Reimers, T., Cooper, L., Wacker, D., Berg, W., Kelly, L., & Allaire, A. (1992). Use of descriptive and experimental analyses to identify the functional properties of aberrant behavior in school settings. *Journal of Applied Behavior Analysis*, *25*, 809–821.

Scott, T. M., Liaupsin, C., Nelson, C. M., & McIntyre, J. (2005). Team-based functional behavior assessment as a proactive public school process: A descriptive analysis of current barriers. *Journal of Behavioral Education*, *14*, 57–71.

Shirley, M. J., Iwata, B. A., & Kahng, S. W. (1999). False-positive maintenance of self-injurious behavior by access to tangible reinforcers. *Journal of Applied Behavior Analysis*, *32*, 201–204.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.

Simpson, R. (2005). Evidence-based practices and students with autism spectrum disorders. *Focus on Autism and Other Developmental Disorders*, *20*, 140–149.

Stanovich, K. E. (2004). *How to think straight about psychology*. New York: Pearson.

Stichter, J., & Conroy, M. A. (2004). A critical analysis of the role of measurement on the validity of emotional and behavioral disorders (EBD) research. *Behavioral Disorders*, *30*, 7–18.

Stichter, J. P., Hudson, S., & Sasso, G. M. (2005). The use of structural analysis to identify setting events in applied settings for students with emotional/behavioral disorders. *Behavioral Disorders*, *30*, 401–418.

Stichter, J. P., Lewis, T. J., Johnson, N., & Trussell, R. (2004). Toward a structural assessment: Analyzing the merits of an assessment tool for a student with E/BD. *Assessment for Effective Intervention*, *30*(1), 25–40.

Tapp, J., & Wehby, J. H. (2000). Observational software for laptop computers and optical bar code readers. In: T. Thompson, D. Felce & F. J. Symons (Eds), *Behavioral observation: Technology and applications in developmental disabilities* (pp. 71–82). Baltimore: Brookes.

U.S. Department of Education. (2002). *No child left behind: A desktop reference*. Washington, D.C.: U.S. Department of Education.

U.S. Office of Special Education and Rehabilitative Services. (2003). *IDEA Final Regulations: Major Issues*. Retrieved January 7, (2004), from http://www.ed.gov/offices/OSERS/Policy/IDEA/MAJOR.DOC

Walker, H. M., Ramsey, E., & Gresham, F. M. (2004). Antisocial behavior in school: Evidence based practices (2nd ed.). Belmont, CA: Wadsworth/Thomson Learning.

Walker, H. M., & Severson, H. H. (1992). *Systematic screening for behavior disorders (SSBD)* (2nd ed.). Longmont: Sopris West.

Weeks, M., & Gaylord-Ross, R. (1981). Task difficulty and aberrant behavior in severely handicapped students. *Journal of Applied Behavior Analysis*, *14*, 449–463.

What Works Clearinghouse (2004, August). Does peer-assisted learning work? Retrieved from www.w-w-c.org.

This page is left intentionally blank

AN EXAMINATION OF SCHOOL-WIDE INTERVENTIONS WITH PRIMARY LEVEL EFFORTS CONDUCTED IN SECONDARY SCHOOLS: METHODOLOGICAL CONSIDERATIONS

Kathleen L. Lane, E. Jemma Robertson and Marona Amandla Leaura Graham-Bailey

ABSTRACT

The issue of school violence and antisocial behavior in public schools is, in fact, one of the most pressing concerns in education today. Schools have responded by designing, implementing, and evaluating multi-level models with progressively more intensive levels of support. The foundation of these models is the primary, or universal, prevention program. To date, most investigations have occurred in elementary schools thereby providing limited insight into intervening in secondary schools. This chapter reviews the literature base of school-wide interventions with primary level efforts conducted in secondary schools with an emphasis on methodological considerations. Content includes the findings of a systematic literature

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 157–199

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19007-2

review, a discussion of quality indicators in relationship to primary prevention efforts, and recommendations for future inquiry.

Today's secondary schools are confronted with a wide range of challenges such as serving an increasingly diverse student body within inclusive environments, teaching progressively more differentiated curricula, adhering to ever more rigorous academic responsibilities, and preparing students to transition into postsecondary environments (e.g., university level instruction, vocational education, and employment; Carter, Lane, Pierson, & Glaeser, 2006; Fuchs & Fuchs, 1994; Lane, Pierson, & Givner, 2004; MacMillan, Gresham, & Forness, 1996). At the same time, secondary schools are asked to manage unprecedented levels of violent and antisocial behavior which has created concerns beyond the school setting for society as a whole (Walker, Ramsey, & Gresham, 2004).

The issue of school violence and antisocial behavior in public schools is, in fact, one of the most pressing concerns in education today. This focus on problem behavior is clearly warranted when we consider that 6% of students report being victims of criminal acts while at school and 28 out of every 1,000 students report being victims of violent crimes within or beyond the school setting (DeVoe et al., 2003). Although the rates of violent crimes perpetrated by youth have declined, the magnitude is still concerning with approximately 2.3 million juveniles being arrested in 2001 and 15% of all violent crimes committed by minors (U.S. Department of Justice, 2001). In addition, there has been an increase in disruptive and insubordinate behaviors exhibited by students. Thus, the incidence of violence is alarming and consequences of violence devastating.

Fortunately, several key pieces of legislation and mandates have been issued in response to school violence and antisocial behavior. Title IV of the *Improving America's Schools Act of 1994*, *the Safe and Drug-Free Schools and Communities Act* (1994) permitted state and local education agencies to design drug and violence prevention plans (Turnbull et al., 2002). This act also prompted the zero tolerance policy for drugs and weapons in that behaviors such as bringing a weapon to school or selling drugs are grounds for immediate suspension or expulsion. This harsh stance is rooted in the belief that these behaviors pose an immediate risk to others and impede the preservation of a safe learning environment. The zero tolerance policy for using or selling drugs and possessing a weapon is also supported in the *Individuals with Disabilities Education Act* (IDEA, 1997).

Shortly thereafter, the White House issued a mandate calling for schools to become nonviolent, safe environments (Dwyer, Osher, & Warger, 1998; Kern & Manz, 2004). The Surgeon General also took action on school violence in the Surgeon General's Report on Youth Violence (2001). This report stated that fewer than 10% of services currently offered by schools and communities to decrease antisocial behavior were evidence-based practices (Satcher, 2001). In an effort to take action against antisocial behavior, the report recommended dismantling antisocial networks, developing positive school climates, increasing academic success, and subscribing to a primary prevention agenda. A similar request was also issued with the reauthorization of IDEA (2004), which called for "providing incentives for whole-school approaches, scientifically based early reading programs, positive behavior interventions and supports, and early intervening services to reduce the need to label children as disabled in order to address the learning and behavioral needs of such children."

Schools have responded to concerns of violence and antisocial behavior, as well, as the mandates discussed above, by shifting away from reactive, punitive approaches to school-wide discipline and moving toward more positive, proactive venues (Horner & Sugai, 2000; Walker et al., 2004). In the past, schools have often relied on reactive disciplinary approaches that involved waiting for a problem to occur and then responding with consequences that typically involved removing students from the school environment (e.g., suspensions and expulsions). More recently, schools have shifted to a proactive, instructional approach to managing behavior that involves clarifying expectations among the faculty, teaching these expectations to all students, providing students opportunities to practice these skills, and reinforcing students for meeting these expectations. Many schools have also provided this universal or primary intervention within the context of a three-tiered model, which contains primary, secondary, and tertiary levels of support.

THREE-TIERED MODELS OF PREVENTION

Three-tiered models of prevention contain three levels: primary, secondary, and tertiary. These level of prevention broaden in intensity with students being identified for more focused interventions using school-wide data. Hence, this type of model is a systematic, data-driven approach designed to meet the needs of all students (Lewis & Sugai, 1999).

The most global prevention effort is *primary prevention* (e.g., school-wide social skills, violence prevention, or conflict resolution programs). This level is designed to prevent harm from occurring by supporting a large number of students who exhibit low levels of at-risk behaviors. Specifically, all students attending a given school participate in the primary intervention plan. The goal is to eliminate circumstances that increase a student's tendency toward developing learning and/or behavior problems. Approximately 80% of students are expected to respond favorably.

Students who are nonresponsive to primary prevention efforts as determined by school-wide data as well as students with moderate risk factors are identified to receive *secondary prevention.* This level of support includes specialized approaches for use with small groups of students with similar behavioral, social, or academic concerns. Secondary interventions include more focused interventions to remediate specific skill or performance deficits and include interventions, such as self-regulation techniques, conflict resolution strategies, or academic instruction in a given area (e.g., reading comprehension). Gresham, Sugai, Horner, Quinn, and McInerney (1998) suggest that 10–15% of students will require secondary supports.

The final level, *tertiary prevention,* is reserved for students who do not respond to secondary efforts and who are exposed to multiple risk factors. Students in need of tertiary levels often have long-standing, complex behavioral problems (Kern & Manz, 2004) which require intensive, ideographic interventions such as function-based interventions (Lane, Umbreit, & Beebe-Frankenberger, 1999), curricular modification, comprehensive intervention that involves other supports (e.g., families and mental health services), or highly intensive academic interventions. Approximately 5–7% of students may require these supports.

As stated by Horner and Sugai (2000), school-wide behavior support "is not a new phenomenon … , but is an approach that is well suited to our times." (p. 231). The foundation of these three-tiered models is the primary prevention plan. While these prevention plans have met with demonstrated success at the elementary level (e.g., Hunter, Elias, & Norris, 2001; Netzel & Eber, 2003; White, Marr, Ellis, Audette, & Algozzine, 2001), less attention has been devoted to examining the efficacy of primary prevention models in secondary schools (e.g., middle, junior high, and high schools). This lack of attention to secondary schools may be due, in part, to the unique challenges that secondary schools pose for researchers and administrators alike.

UNIQUE CHALLENGES OF INTERVENING IN SECONDARY SCHOOLS

Unlike elementary schools where students tend to spend the entire school day in a limited number of settings (i.e., their home classroom and specials), middle and high school students are transient throughout the day, moving from class to class, encountering different peers and teachers at each change. Thus, these adolescents are charged with the task of negotiating the academic and behavioral expectations of their various teachers throughout the day (Isakson & Jarvis, 1999). This is a formidable task given that teachers have varying expectations of the socio-behavioral skills deemed essential for success in school (Lane, Bocian, MacMillian, & Gresham, 2004). Further, many schools lack a unified vision for acceptable student behavior, or, at a minimum, have not employed the procedures necessary to successfully implement school-wide discipline plans (Walker et al., 2004). For example, it may be difficult to identify common sets of behavioral expectations among faculty and staff as the number of people needing to reach consensus increases.

In addition, as students transition from middle to high school, the peer group may yield more influence on student behavior than teachers and administrators (Alspaugh, 1998; Morrison, Robertson, Laurie, & Kelly, 2002). Thus, identifying reinforcing activities or tangibles may become more challenging as reinforcers such as teacher attention and recognition by adults becomes less meaningful.

Finally, many of the discipline problems associated with the adolescent years may look topographically different than those problems observed at elementary and middle school, including covert acts of aggression (e.g., stealing; Loeber, Green, Lahey, Frick, & McBurnett, 2000) and internalizing behavior problems (e.g., eating disorders; Morris, Shah, & Morris, 2002). Similarly, the consequences of aggression become more pronounced as students increase in age and size.

Collectively, these issues as well as the logistics of designing, implementing, and evaluating student outcomes in a more complex school settings (e.g., larger number of students, scheduling difficulties, and greater emphasis on curricular demands) may contribute, in part, to the limited empirical information available regarding best practices for addressing the behavior and academic challenges for high school students. Despite the challenges of implementing school-wide, primary prevention programs in secondary schools, it is imperative that the research and teaching communities identify methods for facilitating implementation of primary prevention programs given the deleterious consequences of school violence and academic underachievement (Lane, Gresham, & O'Shaughnessy, 2002; Lane & Wehby, 2002; Walker et al., 2004).

Purpose

This chapter reviews the literature base of school-wide interventions with primary level efforts conducted in secondary schools with an emphasis on methodological consideration. Content includes the findings of a systematic literature review, a discussion of quality indicators in relationship to primary prevention efforts, and recommendations for future inquiry.

METHOD

Article Selection Procedures

A systematic search of psychology and educational databases (*PsychInfo* and ERIC) was conducted to identify school-wide intervention studies published in journals between 1990 and 2005. Search terms included all possible combinations and derivatives of the following sets of terms: (a) positive, proactive, effective, school-wide, multi-level, multi-tier, or three-tier, (b) behavior or discipline, and (c) support, system, structure, model, or intervention.

To select the articles for review, the first and second authors read and evaluated each title and abstract to determine if the article should be read in entirety to establish if it met inclusion criteria. Next, a master list of journals that published the included articles was developed. Hand searches were conducted for each journal (1990 to present) that published two or more of the included articles to identify any other articles that met inclusion criteria. Searches were conducted in the following journals: *American Educational Research Journal*, *Education and Treatment of Children*, and *Journal of Positive Behavior Interventions.*

Finally, if the articles identified in the above steps referenced a study that might be appropriate for inclusion in this review, the article was retrieved for consideration. Thirty-nine articles were identified as appropriate for further review according to the first author. Next, each article was read in entirety to determine if the article met the following inclusion criteria.

Inclusion Criteria

To be included in this review, each article was required to meet the following criteria: (a) reported a primary level intervention that addressed behavior and/or social skills of all students in the building, (b) implemented the intervention in a secondary school, general education setting, (c) reported student outcomes on behavioral, social, and/or academic variables for the participating schools, and (d) published in a peer-reviewed journal between 1990 and 2005. "Primary level intervention" referred to any universal intervention in which all students in the school participated in the intervention just by virtue of attending school. The primary level intervention may have occurred within the context of a three-tiered model of support (e.g., Gottfredson, Gottfredson, & Hybl, 1993) or may have been a single plan designed only for the entire school without progressively more intensive levels of support (e.g., Shapiro, Burgoon, Welker, & Clough, 2002). Further, the program needed to focus on behavioral or social domains and may or may not have included an academic emphasis as well. Investigations examining district-wide implementation of positive behavior support were excluded (e.g., Nersesian, Todd, Lehmann, & Watson, 2000). Articles focusing on only one or more, but not all, grade levels in a given school were also excluded (Gainer, Webster, & Champion, 1993; Harnett & Dadds, 2004; Hawkins, Catalano, Kosterman, Abbott, & Hill, 1999). Another article by DuRant et al. (1996) was excluded as it was unclear as to whether or not all the students in the two middle schools or just the 20% for whom data were collected participated in the violence prevention curricula as part of their health classes.

Second, all interventions needed to be implemented in a general education, secondary school. This included middle, junior high, and high schools. If the article reported intervention programs implemented in multiple school levels (e.g., elementary and middle schools), the article was included provided that the program description and student outcomes were reported separately for the secondary schools (e.g., Lohrmann-O'Rourke et al., 2000; Shapiro et al., 2002; Sprague et al., 2001; Stevens, De Bourdeaudhuij, & VanOost, 2000). In instances where elementary and secondary school (e.g., middle school) data were reported in aggregate form and it was not possible to separate outcomes for the secondary schools, the article was not included as part of the review (e.g., Pepler, Craig, Ziegler, & Charach, 1994; Rosenberg & Jackman, 2003). Interventions implemented in alternative day schools (e.g., Miller, George, & Fogt, 2005) or in schools for students with intellectual disabilities (e.g., Hetzroni, 2003) were excluded as the purpose of this review was to examine the methodologies and outcomes of primary interventions implemented in general education settings.

Third, the article needed to report student outcomes on behavioral, social, and/or academic variables for the participating schools. Investigations that reported only PBS team members' perceptions of program efficacy and did not include student outcome data were excluded from this review (e.g., Kincaid, Knoster, Harrower, Shannon, & Bustamante, 2002). If the program was described for a secondary school, but student outcome data were not specific to the given school (e.g., Nakasato, 2000), the article was excluded. Further, if the article presented school-wide student outcome data reported in another investigation, the article was also excluded (e.g., Turnbull et al., 2002; Warren et al., 2003) and the referenced article was retrieved. However, one such article, Warren et al. (in press), was not able to be located. Investigations that reported at least one student outcome measure were included in the review, even if the article did not include a full method section. Lastly, in one instance (Skiba & Peterson, 2003), five schools were mentioned in the study; however, data were reported for four schools. Consequently, only those four schools were included in this review.

Finally, articles published in peer-reviewed journals between 1990 and 2005 were included in this review to identify recent studies examining the efficacy of school-wide, primary intervention studies. Dissertation articles and book chapters were excluded as the goal was to draw conclusions based on information that had withstood the test of the peer review process.

Coding Procedures

Articles that met the inclusion criteria were read in entirety by all three authors and coded for the following variables: (a) school characteristics (level (middle, junior high, or high), grades taught, locale (urban, suburban, or rural), size, type (public or private), ethnic constitution, and socio-economic status), (b) purpose, (c) intervention focus and components, (d) research design, statistical analyses, and dependent variables, (e) components related to valid inference making (accuracy of dependent variables, treatment fidelity, social validity, and generalization and maintenance; Lane & Beebe-Frankenberger, 2004; Lane, Beebe-Frankenberger, Lambros, & Pierson, 2001), and (f) intervention outcomes.

RESULTS

Of the 41 articles read, 34% ($n = 14$) met inclusion criteria according to the first authors. The second and third authors each independently read and evaluated half of the 41 articles to assess reliability of the article selection process. There was 100% agreement with the first author. The 14 articles contained data on 63 schools serving secondary students (grades 6–12; see Table 1). This chapter offers a review of the school characteristics, intervention focus and components, research methods, components related to valid inference making, and intervention outcomes for middle and high schools.

School Characteristics

Two articles were published prior to 1997, with the majority of the articles ($n = 12$) published between 1997 and 2005. Of the 63 schools represented in the 14 articles, 6 articles reported treatment-outcomes for 1 school (Kartub, Taylor-Green, March, & Horner, 2000; Lohrmann-O'Rourke et al., 2000; Luiselli, Putnam, & Sunderland, 2002; Mehas et al., 1998b; Taylor-Greene et al., 1997; Taylor-Greene & Kartub, 2000) and one article (Skiba & Peterson, 2003) reported intervention outcomes for four schools without the inclusion of comparison schools. The remaining articles include one (Colvin, Kameenui, & Sugai, 1993, Shapiro et al., 2002) or more (Cook et al., 1999; Gottfredson et al., 1993; Sprague et al., 2001; Stevens et al. 2000) control or comparison schools. Metzler, Biglan, Rusby, and Sprague (2001) included one comparison community that was drawn from two schools (sixth grade students from one school and seventh and eighth grade students from a second school).

All except one article (Skiba & Peterson, 2003) included middle or junior high schools (grades 6–8). Skiba and Peterson reported intervention outcomes of two high schools as well as one junior high and one middle school. One article reported outcomes of secondary students through age 16, but did not explicitly state whether the school was a middle, junior, or high school (Stevens et al., 2000).

The number of students attending each school was reported in all but three articles (Lohrmann-O'Rourke et al., 2000; Mehas, Boling, Sobieniak, Burke, & Hagan, 1998a; Skiba & Peterson, 2003). One article (Shapiro et al., 2002), which reported outcomes of elementary and middle schools, did not report disaggregated enrollment data.

Table 1. School-Wide Interventions with Primary Level Efforts Conducted in Secondary Schools: 1990 to Present.

Author (Journal, Year)	School Characteristics (Level, Number, Grades, Locale, Size, Type, Ethnicity, SES)	Purpose	Intervention and Time Frame	Research Design, Statistical Analysis, and Dependent Variables	Components Related to Valid Inference Making	Outcomes
Colvin et al. (ETC, 1993)	Level: middle; No.: 2 (1 treatment; 1 control); grades: NR; locale: NR: $N = 422$ (control); $N = 449$ (experimental); type: NR; ethnicity: NR; SES: NR; (stated that demographic, physical, and operating features were comparable)	To describe the degree to which a school-wide staff development model (PREPARE), a proactive instructional approach to solving problem behavior using effective staff development procedures influenced problem behavior	School-wide program (PREPARE) implemented using a Teachers-of Teachers model of staff development. *Components* PREPARE: (Proactive, responsive, empirical, and proactive alternatives in regular education) 1. Consistent approach to managing problems 2. School discipline, instrument for student success 3. Positive, preventative strategies 4. Active involvement and support – administrators 5. Commitment to change and participation 6. Teacher-of Teachers staff development model	*Design* Pre-post design with an experimental and control school *Statistical analysis* Visual inspection of office referral data collected during two months (March and April) in during the pre- and post-implementation DV1: number of behaviors per month and per 500 students DV2: percentage of change in the number of consequences pre- to post-	RDV: not mentioned TI: not mentioned SV: T (not mentioned), S (not mentioned) G: not mentioned M: not mentioned	12% Increase in office referral data in the control middle school; 50% decrease in the experimental school. Decreases in consequences (office conferences, suspension, detention, and parent meetings) at the experimental school; increases in "other" and detention consequences at the control school.

Gottfredson et al. (AERJ, 1993)	Level: middle; No.: 8 (6 treatment; 2 control); grades: 6–8; locale: treatment: suburban = 2, urban and suburban = 3, urban = 1; comparison: suburban = 1; urban = 1; treatment: $N = 4{,}513$ (range 460–1,050), comparison $N = 1{,}206$ (490, 716); type: public; ethnicity: % white; treatment: range 3–73, comparison: 0, 70; SES: affluent index (range 1.38–2.51), comparison (1.35, 2.69)	To describe a middle school program designed to decrease inappropriate behavior and report the results of a three-year study to assess outcomes	School-, classroom-, and individual-level to reduce misbehavior *Components* 1. School discipline policy revision 2. Computerized behavior tracking 3. Improved classroom organization and management 4. Positive reinforcement *Strategies* 1. Decrease punitive approaches, increase positive reinforcement 2. Increase clarity of expectations 3. Increase follow-through 4. Improve classroom organization and management Baseline year and 2 years of implementation	*Design* Nonequivalent control group design; pre-post differences for groups High fidelity (H) Medium fidelity (M) Comparison (Low, L). *Statistical analysis* Changes over time, by level of implementation; primarily *t*-tests; effect sizes DV1: classroom environment surveys; teachers and students DV2: teacher ratings of students quarterly; teacher DV3: Effective School Battery; student questionnaire DV4: teacher survey: year end; effectiveness and implementation DV5: school discipline records	RDV: M (for one measure) TI: M, R SV: T (M,R), S (not mentioned) G: not mentioned M: not mentioned	Student reports of classroom organization: M and L implementation schools were similar; slight improvement. Treatment schools improved on student report of classroom order, organization, and rule clarity. Teacher ratings of student attentiveness increased and disruptive behavior decreased in H schools. Ratings of disruptive behavior increased in M schools. Student reports of rebellious behavior increased for all groups; students reported lower levels of punishment in the treatment schools.
Taylor-Greene et al. (JBE, 1997)	Level: middle; No.: 1; grades: 6–8; locale: rural, $N = 530$; type: NR; ethnicity: NR; SES: NR	To examine the impact of a school-wide "opening day" training and on-going behavioral support, on the level o	Opening day and on-going behavioral support Opening day *components*	*Design* Descriptive; pre-post comparison no experimental design	RDV: M, R (entry) TI: not mentioned SV: T (M, R), S (not mentioned) G: not mentioned M: not mentioned	Average number of office referrals decreased from 15 to 8.7 per day (42% reduction). Month-by-month

Table 1. (*Continued*)

Author (Journal, Year)	School Characteristics (Level, Number, Grades, Locale, Size, Type, Ethnicity, SES)	Purpose	Intervention and Time Frame	Research Design, Statistical Analysis, and Dependent Variables	Components Related to Valid Inference Making	Outcomes
		student office referrals across a two-year period	1. Define, teach, and reward the five expectations 2. "High-five" expectations "high-five" (be respectful; be responsible; be there – be ready; follow directions; and hands and feet to self) 3. Explain across six locations (6 lessons 25–30 min each in each setting) Ongoing *components* 1. Provide reminders/precorrection 2. Reward appropriate behavior consistently 3. Provide corrective consequences, booster procedures, targeted support when needed Two years: one baseline; one intervention	*Statistical analysis* Visual inspection of graphs DV1: office referrals per day per month; the total number of referrals per month divided by the total number of school days for that month (average number per day per month). DV2: survey: satisfaction (faculty and staff)		comparison indicated monthly reductions except April. Seventh grades still received the greatest number of referrals, but at 65% of the previous year's level. The most common reasons were (a) repeated minor offenses, (b) defiance, and (c) disruption during baseline year – decrease by 50%. 26 faculty were very satisfied

Mehas et al. (TEC, 1998)	Level: middle; No.: 1; grades: 6–8; locale: NR; size: NR; type: NR; ethnicity: NR; SES: increase in students who were eligible for free or reduced lunch	To describe a school-wide violence prevention program, Second Steps: a violence prevention curriculum, implemented at a middle school and determine if the program was effective in reducing aggressive behavior and increasing prosocial behavior.	Second Step: a violence prevention curriculum *Components* 1. Twenty scripted lessons taught 1–3 times/week during a 45–50 min period 2. Principal, assistant principal, and counselor were coordinators 3. Coordinators supported teachers who needed additional support and prepared materials 4. Time allotted to Second Step implementation, bimonthly staff meetings One year	*Design* Descriptive, post-test only *Statistical analysis* Descriptive interpretation DV1: application of strategy (hypothetical) DV2: number of students reported using a strategy (person life) DV3: anecdotal evidence – number of fights and verbal conflicts DV4: referral data – number of fights and verbal conflicts	RDV: not mentioned TI: M SV: T (M), S (not mentioned) G: not mentioned M: not mentioned	62% wrote a peaceful response to hypothetical conflict that directly reflected content taught. 33% wrote that they used a skill learned. 39% reported a problem and described an appropriate nonviolence response. 16% had not experienced a conflict. 12% wrote a nonpeaceful response Number of fights and verbal conflicts decreased
Cook et al. (AERJ, 1999)	Level: middle; No.: 21; No. of treatment and control schools not listed; grade: 7–8; locale: NR; $N = 12{,}398$; type: public; ethnicity: 66% AA, 24% Caucasian, 4% Asian, and 6% other; SES: 21% free and reduced lunches	To examine whether Comer's School Development Program leads to additive or positive effects on middle school students' outcomes.	Comer's School Development Program *Premise* Student skills can be enhanced by improving relationships and social climate in a school before enhancing the academic focus. Schools establish academic and social goals. The program specifies the processes and structures	*Design* Twenty-one schools were matched on racial composition and achievement test scores and randomly assigned to program or control. *Statistical analysis* Multivariate procedures: ANOVA at school and individual levels; effect sizes	RDV: M, R for some measures TI: M,R SV: T (M,R), S (M, R), P (M, R) G: not mentioned M: M, R.	The randomized experiment analyses Participation in Comer's School Development Program did not influence school climate or student outcomes. The lack of outcomes may have been due to mixed quality of program implementation. Nonexperimental analyses Schools with

Table 1. *(Continued)*

Author (Journal, Year)	School Characteristics (Level, Number, Grades, Locale, Size, Type, Ethnicity, SES)	Purpose	Intervention and Time Frame	Research Design, Statistical Analysis, and Dependent Variables	Components Related to Valid Inference Making	Outcomes
			to establish, monitor, and modify goals. *Structure* 1. School planning and management team 2. Social support team 3. Parent team. Process *Principles* 1. Adult groups cooperate; focus on student needs 2. Schools use a problem-solving orientation 3. Decisions made via consensus Four-year study; multiple 3 – cohorts.	DV1: program implementation measures (survey; parent phone survey; implementation data) DV2: school climate (student and staff) DV3: student moderator and outcomes measures: (e.g., demographics, absenteeism, GPA, and student outcome constructs)		procedures like those specified in Comer's theory may produce desired outcomes. No increases in math scores and may have decreased them.
Kartub et al. (JBPI, 2000)	Level: middle; No.: 1; grades: 6–8; locale: rural; $N = 525$; type: NR; ethnicity: 90% Caucasian; SES: 62% FRL	To address hallway noise levels during lunch periods in a rural middle school using school-wide positive behavior supports.	Reduce hallway noise during transitions to lunch *Components* 1. Quick review of (un)acceptable noise levels during lunch transition (1 day 7-min lesson/period).	*Design* Descriptive, pre-post, nonexperimental (baseline, intervention, and 10-week follow-up for each grade level) *Statistical analysis* Mean score com-	RDV: not mentioned TI: not mentioned SV: T (M), S (not mentioned) G: not mentioned M: M, R	Average decibel levels during baseline were 74.8, 76.5, and 76.8 for grades 6, 7, and 8, respectively. During intervention averages decreased to 67.4, 68.6, and 68.9 for each class. During follow-up (next

			2. Environmental change: dimmed the hall lighting and a small blinking light during lunch transitions. 3. Rewards: 5-min. extra lunchtime for every 3 days with quiet transitions (peer attention) Last two months of the school year.	parison and visual inspection of graphs DV1: median decibel DV2: teacher and hall monitor reports (informal)		school year), new sixth graders mean levels were 67.2 and returning grade 7 and 8 students' mean levels were 67.2 and 67.8. Adults reported reduced noise levels; students were reminding each other to be quiet.
Lohrmann-O'Rourke et al. (JPBI, 2000)	Level: middle; No.: 1; grades: 7–8; locale: rural; size: NR; type: NR; ethnicity: NR; SES: NR	To describe the focus and outcomes of a school-wide PBS program designed to teach the skills necessary to be successful in a diverse world.	Positive behavior support *Core elements* 1. Ongoing team planning 2. Data-based decisions 3. Teaching school rules (full day; instruction in all key areas) 4. Reinforcement of appropriate behavior 5. School-wide policy	*Design* Descriptive *Statistical analysis* One statement regarding reduced level of office referral data DV: Office discipline referrals	RDV: not mentioned TI: not mentioned SV: T (not mentioned), S (not mentioned) G: not mentioned M: not mentioned	30–40% Decreases in office referral data since initial implementation
Stevens et al. (BJEP, 2000)	Level: secondary (middle or high school not specified); No.: 9 (9 elementary 9 secondary – 3 in each condition); grades: NR; locale: NR; $N = 712$ (secondary only); type: NR;	To examine if (a) students involved in bully/victim problems benefit from a school-based anti-bullying program and (b) additional support from the research group is	Flemish anti-bullying program *Modules* 1. Intervention in the school environment (anti-bullying policy, no tolerance of bullying)	*Design* Experimental pre-/post-test comparison including a control. Eighteen schools randomly selected from a pool of 50; randomly assigned to conditions:	RDV: M, R TI: not mentioned SV: T (not mentioned), S (not mentioned) G: not mentioned M: M, R (post 2)	Better outcomes on bullying and victimization among students in the treatment only group compared with students in the treatment + support group. Students in

Table 1. (*Continued*)

Author (Journal, Year)	School Characteristics (Level, Number, Grades, Locale, Size, Type, Ethnicity, SES)	Purpose	Intervention and Time Frame	Research Design, Statistical Analysis, and Dependent Variables	Components Related to Valid Inference Making	Outcomes
	ethnicity: NR; SES: NR	requisite in achieving the desired outcomes.	2. Curriculum-based activities for peer group (social cognitive orientation; four sessions of almost 100 min) 3. Focuses on students directly involved in peer aggression, either as bully or victim (social learning theory)	treatment + support, treatment only, and control *Statistical analysis* 3 (condition) × 2 (education level) × 3 time (pre, post 1, post 2) repeated measures ANOVA, planned contrast comparisons Student self-report Bullying Inventory and the Life in School Checklist DV1: bullying others DV2: being bullied DV3: positive interactions		both intervention conditions did not differ from students in the control groups No significant outcome for condition × educational level × time. Less change in students in the treatment without support.
Taylor-Greene and Kartub (JPBI, 2000)	Type: middle; No.: 1; grades: 6–8; locale: NR; $N = 500$; type: NR; ethnicity: NR; SES: NR	To describe the focus and long-term effects of the PBS High Five program implemented in a middle school.	School-wide behavior program *Components* 1. High Five program 2. Expectations are taught during the first 2 days of the school year; reinforced via token economy	*Design* Descriptive *Statistical analysis* Visual inspection of the total number of office referral data per academic year over a five-year period	RDV: not mentioned TI: not mentioned SV: T (M), S (M) G: not mentioned M: M,R	After 1 year of implementation, ODRs decreased 47%. Five years later, ODRs decreased 68% from initial levels.

			3. School climate committee: plan fall training, monitor outcome data, plan booster and reinforcement activities, communicate with school, and maintain budget.	DV: office discipline referrals (ODRs)		
Metzler et al. (ETC, 2001)	Level: middle; No.: 3; grades: 6–8; locale: NR; $N = 645$ (intervention); $N = 110$, 6th grades from one school; 215, 7th and 8th graders from a second school (comparison); type: NR; intervention school: 92% Caucasian, 5% Hispanic, 2% NA, 1% Asian, 1% AA; comparison 88% Caucasian, 5% NA, 4% Hispanic, 2% Asian, 1% AA; SES: intervention: 42% free, 12% reduced lunch; comparison: 46% free, 6% reduced lunches.	To examine the effects of a consultative approach, school-wide PBS plan to improve behavior management practices in a middle school.	School-wide effective behavior support system, (part of the community builders intervention) *Components* 1. Define rules and expectations 2. Teach expected behaviors 3. Provide praise, rewards for desired behaviors 4. Monitor students' behavior, reinforce rules 5. Use summary data to determine progress and refine intervention plans 6. Intervention school: baseline year, intervention year, and maintenance year	*Design* An A-B design in one school with a comparison community. Pre, post, maintenance. No mention of random assignment *Statistical analysis* Office referrals: interrupted time series analysis. School climate survey: graphic form; comparisons between the same grade levels across years. DV1 Student Report: school climate surveys (praise and awards) DV2: number of tickets DV3: number of good news referrals DV4: number of praise notes DV5: office referrals DV6: student reports of perceived safety	RDV: not mentioned TI: M, R SV: T (M, R), S (not mentioned) G: not mentioned M: M, R	Increases in (a) the proportion of students at the treatment school who reportedly received praise or reward, (b) in Tiger Tickets per year, and (c) good news referrals. Rate of office referrals decreased in the treatment year compared to baseline. A larger proportion of students felt safe at the treatment school. Level of physical and verbal aggression was lower during the implementation year compared to baseline. 79% of teachers said school was safer, 86%—student be-

Table 1. *(Continued)*

Author (Journal, Year)	School Characteristics (Level, Number, Grades, Locale, Size, Type, Ethnicity, SES)	Purpose	Intervention and Time Frame	Research Design, Statistical Analysis, and Dependent Variables	Components Related to Valid Inference Making	Outcomes
				DV7: student reports of being harassed DV8: EBS implementation (social validity and fidelity)		havior had improved. Components implemented with 65–100% treatment integrity.
Sprague et al. (ETC; 2001)	Level: middle; No. of schools: 6 (3 treatment, 3 comparison); grades: 6–8; locale: suburban and urban; *N* treatment: *N* = 1,786, comparison: *N* = 2,085; type: NR; ethnicity: proportion of minority students: average treatment = 6.97; average comparison = 14.9; SES: proportion of free and reduced lunch; treatment average: 39.87; comparison average 37.27	To describe the effects of a universal intervention package aimed at improving the safety and social behavior of students in elementary and middle schools.	Effective behavior support (EBS) and violence prevention program EBS model *Components* 1. Problem and appropriate behaviors are defined 2. Students are taught alternative behaviors 3. Effective incentives and motivational systems 4. Staff commits to the intervention over time 5. staff receives training and feedback 6. monitoring systems 7. Second Steps violence prevention curriculum *Components* 1. Higher order social skills	*Design* Evaluative review. Treatment and comparison schools were chosen by school administrators. Treatment-comparison analysis between 9 treatment (3 MS) and 6 comparison (3 MS) schools. *Statistical Analysis* Visual inspection of graphs; mean scores comparisons DV1: assessing behavior support checklist DV2: Oregon school safety survey: risk and protective factors DV3: school vandalism costs	RDV: no TI: M (R) SV: T (not mentioned), S (not mentioned) G: not mentioned M: not mentioned	Assessing behavioral support in schools: treatment MS: 50% of school wide, 32% common area, 48% classroom, and 30% individual student systems as "in place." Oregon school safety survey: no meaningful differences were detected for treatment and comparison schools. Second Steps Knowledge change: all grade levels in treatment school improved Treatment schools showed reduction in office referral compared to baseline year (−36%) and

			2. Structured, sequenced lessons (over the year) 3. Content: anger management, problem solving, empathy 4. Role-play, integration into regular curriculum. Technical assistance: 1–2 times/month; 20 h of formal training; assistance with problem solving 20–40 h across the year; 8 h inservice (Second Steps); 4 h inservice (EBS)	DV4: Second Step Knowledge Tests DV5: teacher use reports (curriculum) DV6: discipline referrals, attendance, SES, ranking of school; annual; building principal; State Department of Education data base DV7: focus group interviews		greater improvement relative to comparison schools. Focus groups: not presented by elementary and middle schools
Luiselli et al. (JPBI, 2002)	Level: middle; No. 1; grades: 6–8; locale: rural; $N = 623$; type: public; ethnicity: 96.5% European American, 1.4% Hispanic, 1.4% Asian, .6% AA; SES: 7% FRL	To report a longitudinal (four-year) evaluation of a behavior support program implemented with all students in a public middle school.	School-wide intervention with academic and behavioral components *Components* 1. Students received recognition cards – a quarterly lottery (academics, attendance, no detentions/expulsions, or improvement) 2. Three yearly lotteries 3. Caught being good cards (CBG; academic, social, and behavior reasons) 4. CBG weekly drawings	*Design* Descriptive, nonexperimental, over four years Statistical analysis Numerical comparison over four years and visual inspection of graphs DV1: number of detention slips – disruptive-antisocial behavior DV2: number of detention slips – vandalism DV3: number of detention slips – substance use	RDV: not mentioned TI: stated that it was not assessed SV: T (not mentioned), S (not mentioned) G: not mentioned M: M,R	Steady decreases in the number of detention slips for disruptive-antisocial behavior (1,326, 1,237, 717, and 599, respectively) and substance use (9, 6, 6, and 1) over the four years. Overall decreases in vandalism (11, 15, 8, and 5). Stated improvements in attendance and lottery improvements, although the former was not clear from the graph.

Table 1. *(Continued)*

Author (Journal, Year)	School Characteristics (Level, Number, Grades, Locale, Size, Type, Ethnicity, SES)	Purpose	Intervention and Time Frame	Research Design, Statistical Analysis, and Dependent Variables	Components Related to Valid Inference Making	Outcomes
			5. Winners acknowledged (weekly news letter) 6. Detentions (infractions) 7. School adjustment counselor; six special education teachers provided academic support	DV4: percentage of student attendance – ratio of students present on a given day to the number enrolled DV5: percentage of students qualified for lottery drawings		
Shapiro et al. (PITS, 2002)	Level: middle; No.: 4 (3 treatment; 1 control middle); grades: 6–8; locale: urban; N = NR by level (587 including elementary grades 4 and 5); type: public; ethnicity: 88% AA, 8% Caucasian, 1% Hispanic, and 3% other; SES: 32% of students were below poverty, 80% of students received free and reduced lunch	To evaluate the violence prevention effects of The Peacemakers Program with students in fourth through eighth grades.	Violence prevention program: The Peacemakers Program *Components* 1. 17 teacher-led lessons (45 min) 2. Teachers manual 3. Techniques for infusing content into students daily life (12–15 h stated) 4. Teachers prompt the use of techniques, reinforce prosocial behavior, peer reinforcement, and post program materials in hallways and classrooms	*Design* Experimental: pre- and post-program assessments with comparisons of youth who did and did not receive the intervention. Assignment to intervention and control conditions made at the school level (intervention: 3 MS, 3 ES; control 1 MS, 1 ES). Random assignment not stated. *Statistical analysis* 2 × 2 ANCOVAS grade (middle vs. elementary), group (inter-	RDV: mentioned for AGVQ TI: M, R SV: T (not mentioned), S (not mentioned) G: not mentioned M: not mentioned	AGVQ: significantly lower LS means in the younger control group compared to the other three groups, which did not differ from each other. ABC-T: significantly higher LS means in the older control group subjects compared to the three other groups, which did not differ from each other. Suspensions: significant differences between the older control

			5. Remediation component to supplement its primary prevention emphasis (school wide 17 lessons × 45 min)	vention vs. control). Pretest scores: covariates. Teachers measures – sampling procedure: teachers filled out observation scales on the first six students in grade books, (practical approximation of a random sampling procedures) DV1: Attitude Towards Guns and Violence Questionnaire (AGVQ) DV2: knowledge of psychosocial skills DV3: aggressive behavior checklist – S (ABC-S) DV4: aggressive behavior checklist – T (ABC-T) (Teacher) DV5: disciplinary incidents DV6: conflict mediation referrals DV7: suspensions because of violence		group youth and the other groups. Knowledge: ABC-S, disciplinary contacts; conflict mediation: no significant differences
Skiba and Peterson (PSF, 2003)	Level: high, middle, junior; No. 4 schools (2 high, 1 middle, 1 junior high); grades: NR; locale: S1: rural, S2–4: NR; size: NR; type: NR; ethnicity: NR; SES: NR	To describe efforts to increase implementation of effective instructional methods of school discipline and report findings of year-one implementation.	Effective instructional methods of discipline. *Components* S1 (HS) 1. Intervention room 2. Classroom management training workshop	*Design* Descriptive, pre-post, nonexperimental (baseline and year one data) *Statistical analysis* Visual inspection of out-of-school suspension data	RDV: not mentioned TI: mentioned in introduction, not mentioned or reported in findings SV: T (M, informally), S (not mentioned) G: not mentioned M: not mentioned	Out-of-school suspensions declined between 40% and 60%. Gains included students with disabilities with values reported for one middle school.

Table 1. *(Continued)*

Author (Journal, Year)	School Characteristics (Level, Number, Grades, Locale, Size, Type, Ethnicity, SES)	Purpose	Intervention and Time Frame	Research Design, Statistical Analysis, and Dependent Variables	Components Related to Valid Inference Making	Outcomes
			S2 (MS) 1. Safe schools TV show: based on Second Step: violence prevention 2. Parent newsletter: monthly detailing events and activities S3 (HS) 1.Civility themes: school activities and events 2.Alternatives to out-of-school suspension S4 (JH) 1.The code – four principles to guide student behavior (school-wide recognition) 2.Civility curriculum	DV1: out-of-school suspensions DV2: academic success (informal) DV3: social validity: teachers (informal)		Academic gains S1 – received the new American High School Award for reform efforts and increased academic excellence. Social validity: teachers S1 – SRS team members attributed declines in suspensions to the intervention and room and stated that there is no longer of row of student chairs lined up outside the main office.

Note: NR refers to not reported. No. refers to number. Level: MS refers to middle school; JR refers to junior high; HS refers to high school. Ethnicity: AA refers to African American. NA refers to Native American. SES refers to socio-economic status. DV refers to dependent variables. Components: RDV refers to reliability of the dependent variables. TI refers to treatment integrity, SV refers to social validity, G refers to Generalization, M refers maintenance Categories: M refers to mentioned, R refers to reported.

Locales were mentioned in eight articles and reported for 23 schools. Eighteen schools were located in either urban or suburban settings (Gottfredson et al., 1993; Shapiro et al., 2002; Sprague et al., 2001); whereas, only five schools were located in rural settings (Kartub et al., 2000; Lohrmann-O'Rourke et al., 2000; Luiselli et al., 2002; Skiba & Peterson, 2003; Taylor-Greene et al., 1997).

All but three articles (Lohrmann-O'Rourke et al., 2000; Mehas et al., 1998a; Skiba & Peterson, 2003) reported the number of student participants; however, one article (Shapiro et al., 2002) did not report the number of participants separately for elementary and middle school students. Based on reported information, school size ranged from 325 (Metzler et al., 2001) to 1,050 (Gottfredson et al., 1993) students.

The type of institution (public or private) was explicitly stated in four articles (Cook et al., 1999; Gottfredson et al., 1993; Luiselli et al., 2002; Shapiro et al., 2002), with all 34 schools being public. Based on other information provided in the articles (e.g., being a part of a school district), it is likely that the remaining 29 schools were also public schools.

Ethnicity was reported in seven articles, with some schools predominately Caucasian (Kartub et al., 2000; Luiselli et al., 2002; Metzler et al., 2001) and others predominantly African American (Cook et al., 1999; Shapiro et al., 2002).

The majority ($n = 8$) of articles provided data regarding socio-economic status (e.g., percentage of free and reduced lunch, poverty levels (Shapiro et al., 2002), or affluence index (Gottfredson et al., 1993)). The percentage of students receiving free or reduced lunch ranged from a low of 7% (Luiselli et al., 2002) to a high of 80% (Shapiro et al., 2002).

Interventions: Focus and Components

The focus of most school-wide interventions was primarily behavior. Six studies explicitly discussed effective behavior supports and positive behaviors supports focused on either the entire school as a setting (Lohrmann-O'Rourke et al., 2000; Metzler et al., 2001; Sprague et al., 2001; Taylor-Greene et al., 1997; Taylor-Greene & Kartub, 2000) or a specific setting (e.g., hallways, Kartub et al., 2000), four studies examined violence prevention (Mehas et al., 1998a; Shapiro et al., 2002; Sprague et al., 2001) or anti-bullying (Stevens et al., 2000) programs, two examined instructional models of school-wide discipline (Gottfredson et al., 1993; Skiba & Peterson, 2003), one focused on using a teachers-of-teachers model to implement

Project PREPARE (proactive, responsive, empirical, and proactive alternatives in regular education) (Colvin, Kameenui, & Sugai, 1993), and one focused on improving interpersonal relationships and social climate as art of the Comer's School Development Program (Cook et al., 1999). One study addressed both behavioral and academic domains in the school-wide intervention (Luiselli et al., 2002).

Each study provided core-components of intervention program which included clarification, instruction, and reinforcement for meeting expectations (e.g., Kartub et al., 2000; Lohrmann-O'Rourke et al., 2000; Luiselli et al., 2002; Metzler et al., 2001; Shapiro et al., 2002; Skiba & Peterson, 2003; Sprague et al., 2001; Taylor-Greene et al., 1997; Taylor-Greene & Kartub, 2000). Some articles reported intervention intensity, or dosage, when the intervention included a packaged intervention curriculum such as the Peacemakers Program (violence prevention, Shapiro et al., 2002), Second Step: A Violence Prevention Curriculum (Mehas et al., 1998a), and the Flemish Anti-bullying Program (Stevens et al., 2000). Dosage was also provided for the initial methods used to teach students the expectations (e.g., Kartub et al., 2000; Taylor-Greene et al., 1997). Other articles provided specific information regarding the level of support provided to prepare the school site for implementation (Sprague et al., 2001).

Research Design, Statistical Analyses, and Dependent Variables

Research design and statistical analyses. Studies that involved just one school were descriptive, nonexperimental in nature using pre-post comparisons in one school (Kartub et al., 2000; Lohrmann-O'Rourke et al., 2000; Luiselli et al., 2002; Mehas et al., 1998a; Taylor-Greene et al., 1997; Taylor-Greene & Kartub, 2000), most of which included graphs that could be analyzed visually. Kartub and colleagues included a 10-week follow-up assessment point for each grade level. Luiselli et al. (2002) studied implementation over a four-year period and Taylor-Greene and Kartub over a five-year period.

Other studies involved comparison schools in their design. Colvin et al. (1993) employed a pre-post design with an experimental and control school. However, it was not possible to analyze data using school as the unit of analysis given that there was only one school in each condition. Data were analyzed using visual inspection (Colvin et al. 1993). Metzler et al. (2001) employed an AB design in one school with a comparison community drawn from two schools (sixth graders from one school and seventh and eighth grade students from a second school) assessed at three time points (pre-, post-, and maintenance). Office referral data were analyzed with interrupted time series analysis (ITSACORR) and school climate survey data were analyzed visually with comparisons made between the same grade levels across years. The study conducted by Sprague et al. (2001), although not a true experiment, conducted an evaluative review of 15 schools (9 elementary and 6 middle schools), with 3 middle schools receiving the treatment and 3 serving as comparison schools. The intervention schools participated in the intervention concurrently and data were collected from treatment and comparison schools on the same schedule. The study was not a true experiment given that schools were not randomly selected and were chosen by school administrators (Sprague et al., 2001). Data were analyzed via visual inspection of graphs and mean score comparisons. Further, Gottfredson et al. (1993) reported using a nonequivalent control group design with high, medium, and low (comparison) schools assessed pre- and post-implementation analyzed primarily using *t*-tests and effect sizes comparisons.

Other investigations employed more rigorous evaluations by moving toward experimental designs. Shapiro et al. (2002) compared examined the effects of the Peacemakers Program with elementary and middle school students. They stated that an experimental, pre-post design was employed to compare youth who did and did not receive the intervention with assignment made at the school level (treatment: 3 elementary and 3 middle schools; control: 1 elementary and 1 middle school). However, random assignment of schools was not stated. Data were analyzed using multivariate procedures with 2 (grade: elementary, middle) × 2 (group: intervention, control) analysis of covariance (ANCOVAS) using pre-test scores as a covariate given that there were initial differences between the intervention and control groups. A sampling procedure was used for teacher measures in which the teachers completed observations scales on the first six students listed in their grade books. This was done to serve as a practical approximation of random sampling procedures. Stevens et al. (2000) reported conducting an experimental, pre-post test comparison including a control. Eighteen schools (9 elementary and 9 secondary) were randomly selected from a pool of 50 schools and then randomly assigned to one of three conditions: treatment with support, treatment only, and control. Data were analyzed using a 3 (condition) × 2 (education level: elementary, secondary) × 3 (time: pre, post 1, post 2) repeated measures analysis of variance (ANOVA) with planned contrast comparisons. Finally, Cook et al.'s (1999) study involved 21 schools (excluding two pilot schools) which were match on racial composition and achievement test scores for the two prior years. The schools were assigned randomly to treatment or control conditions. Data were analyzed at school and individual levels using multivariate procedures (ANOVA), with the former being the most appropriate method of analysis (Cook et al., 1999). Effect sizes were also computed and reported to examine the magnitude of change.

Dependent variables. Dependent variables included a range of measures to assess student behavior and academic performance; school climate; program implementation; reinforcement; as well as student and teacher perceptions. All but two studies (Cook et al., 1999; Stevens et al., 2000) incorporated office discipline referral, detention (Luiselli et al., 2002), or suspension (Shapiro et al., 2002; Skiba & Peterson, 2003) measures. Two studies include office discipline referrals as the only outcome measure (Lohrmann-O'Rourke et al., 2000; Taylor-Greene & Kartub, 2000). Three studies measured school or classroom climate from teacher and/or student perspectives (Cook et al., 1999; Gottfredson et al., 1993; Metzler et al., 2001).

Program implementation was a dependent variable in four studies and was assessed in a variety of methods including team or teacher surveys or reports (Cook et al., 1999; Gottfredson et al., 1993; Sprague et al., 2001), parent phone interviews (Cook et al., 1999), and process measures of EBS implementation (Metzler et al., 2001). The extent to which students received reinforcement components specified in the intervention programs was assessed in three studies. Specific dependent measures included referrals to the office for positive behaviors (Gottfredson et al., 1993); tickets, praise notes, and good news referrals (Metzler et al., 2001); and percentage of students who qualified for lottery drawings (Luiselli et al., 2002).

Dependent measures also included student perceptions of or attitudes toward school safety (students; Metzler et al., 2001; Sprague et al., 2001), harassment (Metzler et al., 2001), aggressive behavior (Shapiro et al., 2002), guns and violence (Shapiro et al., 2002), and either being bullied or acting as a bully (Stevens et al., 2000). In addition, one study assessed students' knowledge (Second Steps Knowledge Tests about violence prevention, pre- and post-instruction; Sprague et al., 2001) and another assessed students' knowledge of psychosocial skills (Shapiro et al., 2002). Mehas et al. (1998a) also assessed the extent to which students used the strategies taught as part of Second Steps: a Violence Prevention Program in hypothetical and actual situations. Finally, dependent measures were also included to obtain teacher ratings of students' behavior (Gottfredson et al., 1993; Shapiro et al., 2002).

Components Related to Valid Inference Making

If intervention studies are to lead to accurate inferences regarding student outcomes, four components must be included: reliability of the dependent variables, fidelity of intervention implementation (the independent variable; Gresham, 1989; Lane et al., 2004), social validity (goals, procedures, and outcomes; Kazdin, 1977; Wolf, 1978), and generalization and maintenance (Lane & Beebe-Frankenberger, 2004). Nine studies (64%) neither mentioned nor reported reliability of the dependent variables. Two studies mentioned, but did not report, reliability of one of the dependent variables (Gottfredson et al., 1993; Shapiro et al., 2002). Cook et al. (1999) mentioned and reported some reliability information, but, not for all measures. Stevens et al. (2000) mentioned and reported reliability information as did Taylor-Greene et al. (1997, entry of office referral data).

Even fewer studies ($n = 6$; 43%) neither mentioned nor reported treatment integrity. One study stated that treatment integrity was not assessed and was dealt with as a limitation (Luiselli et al., 2002). One study mentioned treatment integrity in the introduction, but did not mention or report treatment integrity in the method or results (Skiba & Peterson, 2003). Mehas et al. (1998a) mentioned treatment integrity, but did not report levels of implementation. The remaining articles ($n = 5$; 36%) mentioned and reported levels of treatment implementation (Cook et al., 1999; Gottfredson et al., 1993; Metzler et al., 2001; Shapiro et al., 2002; Sprague et al., 2001).

Eight studies (57%) mentioned social validity from the teacher ($n = 8$), student ($n = 2$), and/or parent perspectives ($n = 1$), with Cook et al. (1999) assessing and reporting social validity data from all three perspectives. Of the eight studies that mentioned social validity, levels of fidelity were reported (formally or informally) in all but three articles (Kartub et al., 2000; Mehas et al., 1998a; Taylor-Greene & Kartub, 2000).

Six studies (43%) mentioned and reported maintenance data for their school-wide primary plans (Cook et al., 1999; Kartub et al., 2000; Luiselli et al., 2002; Metzler et al., 2001; Stevens et al., 2000; Taylor-Greene & Kartub, 2000). None of the studies mentioned or reported generalization data.

Outcomes

Although the number of primary interventions was rather limited ($n = 14$), outcomes were generally favorable. While a detailed discussion of each study is beyond the scope of this chapter, an overview of treatment-outcomes is presented below.

In general, results of school-wide interventions focusing on effective behavior and positive behavior supports were favorable resulting in decreases in office referral data over time (Colvin et al., 1993; Lohrmann-O'Rourke et al., 2000; Metzler et al., 2001; Sprague et al., 2001; Taylor-Greene et al., 1997; Taylor-Greene & Kartub, 2000) and lower levels of physical and verbal aggression (Metzler et al., 2001). In addition, these interventions yielded increases in recognition (e.g., praise and awards) for meeting expectations (Metzler et al., 2001), and, in some instances (e.g., Metzler et al., 2001), safety. However, improvements in safety were not observed in all investigations (e.g., Sprague et al., 2001). Fidelity data suggest that components were implemented with moderate to high levels of integrity (65–100%; Metzler et al., 2001); however, only one of these studies reported treatment fidelity data.

Studies focusing on violence prevention or anti-bullying were mixed. Outcomes from Shapiro et al. (2002) intervention yielded no significant differences in disciplinary incidences, knowledge acquired, student reports of aggressive behavior, or attitudes toward guns and violence. However, decreases in suspensions and teacher reports of aggressive behavior were observed for middle school students relative to elementary students (Shapiro et al. (2002)). Similarly, outcomes from Stevens et al.'s (2000) study indicated that neither of the two treatment conditions (treatment only or treatment with support) outperformed the control condition in terms of bullying others or being bullied. However, students in the treatment only condition showed better outcomes on these measures as compared to students in the treatment plus support condition. This may have been due to the fact that students in the treatment with support group showed lower levels of bullying and victimization compared to the treatment only condition at the program onset. In contrast, Mehas et al. (1998a) study of Second Steps: a Violence Prevention Curriculum produced favorable outcomes as evidenced by (a) the majority (62%) of students citing peaceful responses to hypothetical conflicts, (b) 72% of students either using a skill from their Second Steps program or another peaceful strategy to resolve an actual conflict, and (c) decreases in the number of fights and verbal conflicts.

Programs focusing on instructional models of school-wide discipline also produced mixed results. Gottfredson et al.'s (1993) examination of high, moderate, and low (comparison) implementation schools produced improvements for both treatment groups on classroom order, classroom organization, and rule clarity from the student perspective with high implementation school producing the greatest gains. While teacher reports of attentiveness increased and disruptive behavior decreased in high implementation schools, increases in disruptive behavior were observed in moderate implementation schools. None of these improvements was observed in low implementation schools. In contrast, Skiba and Peterson (2003) reported 40–60% declines in out-of-school suspensions which included improvements for students with disabilities. However, treatment fidelity data were not reported for the later study.

The remaining studies also produced some positive outcomes including: (a) reduced levels of hallway noise as measured by median decibel levels across grade levels which were maintained into follow-up (Kartub et al., 2000); and (b) decreases in office referrals and consequences such as office conferences, suspensions, detentions, and parent meetings (Colvin et al., 1993). The Comer's School Development Program (Cook et al., 1999), however, did not influence student outcomes or school climate. This lack of findings may have been due to varied levels of intervention fidelity. Non-experimental analysis suggested that schools using procedures such as those specified in Comer's theory may be associated with positive changes in social behavior and adjustment. Finally, the one study that addressed both behavioral and academic domains in the school-wide intervention (Luiselli et al., 2002) produced steady declines in the number of detention slips issued for disruptive-antisocial behavior and substance abuse over the four-year period. They also cited improvements; however, this was not clear from the data presented in the graph.

In sum, findings from these studies can best be described as cautiously optimistic. The vast majority of these investigations contain methodological concerns that potentially influence the ability to interpret validly intervention outcomes. Limitations of these studies and the corresponding effects on interpretations of the present studies and directions for future investigations will be offered in the following discussion section.

Limitations

As evidenced above, the body of research examining the efficacy of school-wide primary interventions at the secondary level is limited. This is concerning given that primary interventions have the potential to prevent harm for a large number of students and require relatively fewer resources as compared to secondary and tertiary level of support (Lane, Gresham, & O'Shaughnessy, 2002; Lewis & Sugai, 1999; Walker & Severson, 2002).

The majority of these studies focused on middle schools with only one study (Skiba & Peterson, 2003) explicitly reporting on a high school.

Although many of the investigations produced desirable outcomes (e.g., lower rates of office referrals), a number of methodological limitations limit the ability to draw accurate conclusions about intervention outcomes. Namely, some of these limitations include: (a) a scarcity of investigations at the high school level; (b) a limited number of investigations of schools located in rural regions; (c) limited demographic information provided on the participating schools; (d) intervention descriptions insufficient to allow replication; (e) research designs that were predominately descriptive thereby preventing causal conclusions to be drawn; (f) limited scope of outcome measures, often described without reliability and validity information; and (g) designs that lack the components required to draw valid conclusions about intervention outcomes (Lane et al., 2001). These limitations will be discussed within the context of quality indicators for intervention research and directions for future investigations of school-wide primary interventions at the secondary level will be discussed.

DISCUSSION: FINDINGS, CONCERNS, AND RECOMMENDATIONS

During the past decade, policies such as Title IV of the Improving America's Schools Act of 1994, the Safe and Drug-Free Schools and Communities Act (1994) and the reauthorization of the IDEA (1997) have prompted a focus on violence and drug prevention programs, a zero tolerance stance on drugs and guns, and an emphasis on positive behavior supports. Specifically, the Safe and Drug-Free Schools and Communities Act (1994) enabled state and local agencies to develop violence and drug prevention programs and prompted the zero tolerance policy for students possessing or using drugs or possessing weapons. IDEA also supported the zero tolerance policy for drugs and alcohol and at the same time mandated positive behavior supports. Further, the Surgeon General's (2001) report listed antisocial behavior in schools as a primary concern. This report recommended the following responses: dismantling antisocial networks, increasing academic success, developing positive school climates, and adopting a primary prevention agenda, all of which can be addressed within the context of school-wide, universal supports. Collectively, these policies have set the stage to focus on schools as an agent for change (Sugai & Horner, 2002a, b).

This chapter reviewed the literature on school-wide interventions with primary level efforts conducted in secondary schools. The studies conducted to date, although limited in number, largely descriptive in nature, and noted for certain methodological concerns, have provided important direction on how to develop this line of inquiry.

Many of the descriptive studies described in this review present student data that illustrates decreases in problematic behavior over time (e.g., Colvin et al., 1993; Lohrmann-O'Rourke et al., 2000; Metzler et al., 2001; Sprague et al., 2001; Taylor-Greene et al., 1997; Taylor–Greene & Kartub, 2000). Yet, several of these investigations do not incorporate core components (e.g., reliability of the dependent variables, fidelity of implementation of the independent variable, social validity, or generalization and maintenance data) necessary to draw valid conclusions. Nor, do they employ research designs that will allow causal conclusions to be drawn. While descriptive studies provide important information to inform future intervention efforts, it is important to move beyond descriptive investigations toward more rigorous, experimental or quasi-experimental designs (Kern & Manz, 2004).

If the goal of all intervention research is to produce meaningful lasting changes (Baer, Wolf, & Risely, 1968) and to be able to draw definitive, causal conclusions about the relationship between the intervention and student performance, then it is essential that studies employ experimental, or even quasi-experimental research designs. Although randomized trials are not often used in educational research, such studies represent the "gold-standard" in that randomized trials are the best method for identifying systematic relationships between independent (e.g., interventions) and dependent variables (e.g., student outcomes, Feurer, Towne, & Shavelson, 2002). In addition, not only is it important that these designs be utilized, but also that they are constructed using high quality standards. While this often proves to be a formidable challenge when conducting school-based treatment-outcome studies, particularly at the school level; this is a challenge we as a field must accept.

Gersten, Fuchs, Compton, Coyne, Greenwood, and Innocenti (2005) have delineated stringent quality indicators for group experimental and quasi-experimental research in special education (see Table 2). To be considered a study of acceptable quality, a study would meet all except one essential quality indicators and incorporate a minimum of one desirable quality indicator. To be considered a high quality study, a study would again incorporate all except one essential quality indicator and include at least four desirable quality indicators. We will briefly review both the quality indicators recommended by Gersten et al. and discuss how these indicators have been, or need to be, addressed in the literature focusing on school-wide primary intervention efforts at the secondary level.

Table 2. Essential and Desirable Quality Indicators for Group Experimental and Quasi-Experimental Research Articles and Reports.

Essential Quality Indicators

Quality indicators for describing participants

1. Was sufficient information provided to determine/confirm whether the participants demonstrated the disability(ies) or difficulties presented? Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?
2. Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

Quality indicators for implementation of the intervention and description of comparison conditions

1. Was the intervention clearly described and specified?
2. Was the fidelity of implementation described and assessed?
3. Was the nature of services provided in comparison conditions described?

Quality indicators for outcome measures

1. Were multiple measure used to provide an appropriate balance between measures closely aligned with the intervention[a] and measures of generalized performance?
2. Were outcomes for capturing the intervention's effect measured at the appropriate times?

Quality indicators for data analysis

1. Were die data analysis techniques appropriately linked, to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?
2. Did the research report include not only inferential statistics but also effect size calculations?

Desirable Quality Indicators

1. Was data available on attrition rates among intervention samples? Was severe overall attrition documented? If so, is attrition comparable across samples? Is overall attrition less than 30%?
2. Did the study provide not only internal consistency reliability but also test–retest reliability and interraer reliability (when appropriate) for outcome measures? Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?
3. Were outcomes lot capturing the intervention's effect measured beyond an immediate posttest?
4. Was evidence of the criterion-related validity and construct validity of the measures provided?
5. Did the research team assess not only surface features of fidelity implementation (e.g., number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?
6. Was any documentation of the nature of instruction or series provided in comparison conditions?
7. Did the research report include actual audio or videotape excerpts that capture the nature of the intervention?
8. Were results presented in a clear, coherent fashion?

[a] A study would be acceptable if it included only measures of generalized performance. It would not be acceptable if it only included measures that are rightly aligned.

Reprinted with permission from Gersten et al. (2005).

Conceptualization

Gersten et al. (2005) offer four indicators for determining the extent to which a study establishes the context and design of a study. First, the study is either (a) designed based on the results of seminal, related investigations that lead to the purpose of the study or (b) represents a unique approach that is grounded in research and sound conceptualization. Second, a compelling case is made for the significance of the research study. Third, a case must be made to justify the proposed intervention and explain the experiences held by the comparison groups. Finally, the research questions are clearly stated and appropriate to the intent of the study.

In light of public concern surrounding school violence, current legislation (e.g., IDEA, 1997; *the Safe and Drug-Free Schools and Communities Act*, 1994; Surgeon General's Recommendations, 2001), the eroding "capacity of our society to safely raise and socialize our children" (Walker, 2003), it is more than possible to establish a context for this work and build a compelling case for conducting randomized trials of school-wide, primary level interventions. This line of inquiry will be able to establish salience validity and incidence validity (Fabes, Matrin, Hanish, & Updegraff, 2000). It incorporates salience validity in the sense that the educators as well as general public are aware of the magnitude and consequences of school violence. It also incorporates incidence validity in that primary interventions impact a large number of students – namely, the entire student body.

While some of the studies described in this review have included comparison schools (e.g., Colvin et al., 1993; Cook et al., 1999; Gottfredson et al., 1993; Metzler et al., 2001; Shapiro et al., 2002; Sprague et al., 2001; Stevens et al., 2000), many descriptions of school-wide practices for comparison sites are insufficient to provide a clear explanation of the focus, scope, and magnitude of the school-wide intervention used in the comparison schools. Some studies (e.g., Gottfredson et al., 1993) have attempted more sophisticated designs that contrasted schools with high, moderate, and low (comparison) levels of fidelity. Still other studies (e.g., Cook et al., 1999) have used information about the comparison site characteristics to interpret outcomes. Yet, overall, there is a lack of clarity on the experiences occurring in the comparison schools.

Finally, while the clear majority of the articles included a statement regarding the purpose of the study; none of studies that included comparison groups provided explicitly stated research questions that were tied to the purpose statement. Moving forward, this line of inquiry could be enhanced by affording greater attention to the comparison schools and providing explicitly stated research questions, linked to the study's purpose.

Participants and Sampling

Gersten et al. (2005) also suggest the following indicators for describing participants. First, essential information is provided to judge the accuracy of the disability(ies) category(ies) under investigation and guide decisions about to who the findings may generalized. Second, procedures such as random assignment are used to increase the likelihood that participants in different conditions are comparable at the onset of the study. Third, attrition is documented and explored to examine the similarity of attrition rates across groups. Fourth, essential information is provided to determine if the intervention providers for different conditions are comparable across groups.

A few studies included in this review provided comprehensive information regarding the participants including information such as grade level, locale, the type of institution (public or private), ethnicity, and socio-economic information (e.g., Gottfredson et al., 1993; Luiselli et al., 2002; Sprague et al., 2001). However, several studies did not provide explicit information regarding school demographic that limits external validity of the findings. Future investigations should provide detailed information regarding school characteristics to address this limitation. Further, future studies would be wise to examine how different types of students (e.g., those with internalizing and externalizing behavior patterns; documented and verified disabilities; gifted students; and typical students) respond to school-wide primary intervention efforts. The assumption in many studies is that all students respond uniformly to primary level supports. However, it is quite possible that all students are not participating equally in the intervention program (e.g., equal levels of reinforcement) and consequently, may not respond uniformly to primary level interventions. The field needs to move beyond the question of "How did the school as a whole respond to the primary level plan?" and extend this line of inquiry to determine "How do different types of students respond to the primary level plan?"

Schools were assigned to either treatment or comparison conditions in three studies (Cook et al., 1999; Shapiro et al., 2002; Stevens et al., 2000); however, randomly assignment was explicitly stated in two studies (Cook et al., 1999; Stevens et al., 2000). Cook et al. (1999) matched 21 schools on racial composition and achievement test scores two years prior to intervention onset and then randomly assigned these schools to either the treatment or control conditions. In other instances, treatment and comparisons schools were selected by school administrators (Gottfredson et al., 1993; Sprague et al., 2001) and the lack of a formal experimental comparison was acknowledged in one of these studies (e.g., Sprague et al., 2001). In other cases there was no mention of how schools were assigned to conditions other than the fact that assignment to intervention and control conditions was made at the school level (e.g., Colvin et al., 1993; Metzler et al., 2001; Shapiro et al., 2002). In school-wide, primary interventions, school is the unit of analysis and, consequently, random assignment should occur at the school level. However, random assignment does not guarantee equivalent study groups (Gersten et al., 2005). Therefore, it may be wise to match schools on key characteristics (e.g., socio-economic status, size, and region) and then randomly assign one school from each pair to each condition or conduct a stratified random assignment procedure as was attempted in Cook et al. (1999) study. We do recognize that conducting experimental studies with schools as the unit of analysis results in a very large-scale study and dramatically increases the cost and scope of the study. Yet, if we are truly committed to drawing causal conclusions about intervention outcomes, this is the necessary venue.

Similarly, only four studies dealt with issues of attrition (Luiselli et al., 2002; Metzler et al., 2001; Shapiro et al., 2002; Stevens et al., 2000). Although a complex task, particularly when working with secondary-level schools, greater attention must be given to reporting attrition and mobility. For example, if a school is characterized by high mobility rates, then data should either (a) be analyzed separately for students who were at the school for the entire year given that the intervention dosage (treatment length) vary for students who were and were not present for the entire year or (b) the data analysis procedures should incorporate the dosage into the design (e.g., covariate) to control for initial levels of variability.

Finally, information on teacher characteristics such as credentialing status, years of teaching experiences, educational attainment, perceptions of social validity at program onset, and fidelity of implementation should be used to predict student outcomes. These data should be used to determine if teachers (who are the interventionists in school-wide, primary interventions) in intervention and control schools are comparable at program onset. This is necessary to remove teacher differences as a possible confound when interpreting intervention outcomes.

Implementation of the Intervention and Features of the Comparison Condition

Gersten et al. (2005) offer three quality indicators for intervention implementation and comparison conditions. First, the intervention procedures must be described with clarity and sufficient detail to allow replication.

Second, the presence and quality of treatment integrity should be reported. Third, the educational practices occurring in the comparisons condition is provided and documented.

Precise descriptions of intervention procedures are essential not only for replication, but also for determining how practices converge and diverge across studies. Consequently, interventions must be described with precisions. While many studies provided clear precise description of the general intervention procedures, often times there was insufficient detail to allow for replication (e.g., Cook et al., 1999). One component that requires particular attention in future research is the reinforcement component. Namely, are all adults afforded the opportunities to deliver "tickets"? Is there a limit on how many tickets adults can allocate? Is there documentation as to which students receive tickets? Is there documentation to determine if students who receive tickets enter these tickets in the lotteries? In brief, if the goal is replication, intervention descriptions will require a greater degree of specificity.

As previously mentioned, only five studies reported implementation fidelity (Cook et al., 1999; Gottfredson et al., 1993; Metzler et al., 2001; Shapiro et al., 2002; Sprague et al., 2001). It is critical that future studies not only describe the presence or absence of each intervention component, but also that the quality of implementation be documented as well. Further, this information should be used when analyzing student outcomes as was done in Gottfredson et al. (1993) study. The absence of treatment integrity data poses clear threats to the internal and external validity of any study. Given the costs associated with conducting school-wide, primary interventions – treatment integrity data are mandatory.

Finally, data must be collected on comparison schools to determine (a) the nature of the school-wide policies and practices and (b) the extent to which similar practices are occurring in the intervention and comparison sites. When treatment contamination occurs, this has direct implications for interpreting intervention outcomes. For example, in the study by Cook et al. (1999), the lack of treatment effects may have been due to the educational practices that took place in the comparison schools. Namely, the comparison schools may have contained practices that were similar to the treatment school's practices.

Outcome Measures

Gersten et al. (2005) offer five quality indicators pertaining to outcome measures. First, multiple measures are used to provide a balance between proximal and more distal measures of performance. Second, information regarding the reliability and validity of measures are reported. Third, measures are administered at the appropriate times necessary to examine the intervention impacts. Fourth, data collectors and scorers are blind to study conditions and participants have parallel familiarity with data collectors. Fifth, interscorer agreement is reported with reliability in data collection and scoring at 0.90 or above.

Too often the attention afforded to intervention design is not afforded to selecting measures to evaluate intervention outcomes. If the goal of school-wide research to reduce problem behaviors, increase adaptive behaviors, and improve academic performance, then it is essential to have multiple measures for each domain of interest with some measures assessing proximal effects and other assessing more distal effects. Too often studies in school-wide primary interventions have relied exclusively on one outcome measure – often reported without reliability information – to evaluate intervention outcomes (Kern & Manz, 2004).

Further, it is important that these measures be psychometrically sound, including reports of alpha coefficients (estimates of internal consistency reliabilities) and indications of concurrent validity. Yet, as evidenced in this review, this information is not reported often, with nine studies neither mentioning nor reporting reliability of the dependent measures. Without reliability information, one cannot be certain that changes in the outcome measures are due to the intervention or due to measurement error (Kleinbaum, Kupper, Muller, & Nizam, 1998).

In terms of the timing of assessments, school-wide, primary interventions have often collected baseline data, intervention data, and in some cases (e.g., Cook et al., 1999; Kartub et al., 2000; Luiselli et al., 2002; Metzler et al., 2001; Stevens et al., 2000; Taylor-Greene et al., 2000) maintenance data. However, designs could be improved and more accurate conclusions drawn about sustainability and capacity building at the site level by collecting maintenance and generalization data over a longer period of time. Change at the school-level often takes a few years to occur. Therefore, if the goal is meaningful, lasting change, then assessments will need to be taken over the course of several years. Admittedly, this poses fiscal, personnel, and other resource challenges given that (a) most federally funded projects span between three and five years making long-term follow-up difficult, (b) schools, from our experience, often reach a point where they want to implement the program with less university support and greater independence, and (c) the logistics of data collection become more difficult when working in rural settings – an understudied locale.

Although it would be ideal to have data collectors and scorers are blind to study conditions, this is often not possible when implementing school-wide, primary level interventions given that implementation evidence is visible within the building (e.g., posters depicting expectations, reward assemblies, and instructional videos). However, all data collectors need to be trained to criterion and these training procedures and reliability standards need to be reported in the method section of all studies.

Although this indicator increases the costs of conducting school-based interventions, it is imperative that data collection and scoring procedures be specified, taught, and held to a high standard of reliability (0.90). Again, documentation of reliability of administration and scoring is essential to draw accurate conclusions about intervention outcomes. To this list we would also encourage future investigations to report accuracy of data entry (25%; e.g., Taylor-Greene et al., 1997), particularly when investigations report a narrow range of outcome measures. It is alarming to see how few studies in this body of literature have met any of these essential quality indicators of outcome measures.

Data Analysis

The final set of essential quality indicators pertain to data analysis. Gersten et al. (2005) offer several indicators emphasizing (a) a clear connection between the design and unit of analysis to the primary research question and (b) statistical analysis techniques that take initial differences between groups into account. They highlight these themes by offering the following recommendations. First, use multilevel analyses when possible as they are "designed to consider multiple units within a single analysis" (Gersten et al., p. 161). Second, recognize that the unit analysis may be different for different measures within the same study. Third, select and justify data analysis techniques in relationship to the research question and hypothesis. Fourth, use sampling or statistical techniques (e.g., ANCOVAS) to account for variability in a sample. Fifth, the unit of analysis should be linked to the main statistical analyses. Sixth, a power analysis should be conducted for each unit of analysis under investigation to determine minimum cell size.

This set of quality indicators is one of the most challenging tasks for researchers interested in conducting school-wide, primary interventions. In most cases the school will serve as the unit of analysis that necessarily expands the scope of treatment-outcome studies. Some studies included in this review have employed multivariate analyses that included analyses at the school level and one study explicitly stated the unit of analysis in relationship to the method of analysis (e.g., Cook et al., 1999). However, this is clearly not the standard. It is particularly difficult to incorporate school as the unit of analysis and when exploring implementation in schools with low incidence occurrences (e.g., rural schools). Yet, additional efforts are needed to describe the efficacy of school-wide, primary interventions conducted in urban, suburban, and rural areas given that schools in these different locales are likely to vary in terms of available resources, level of risk, and philosophy regarding positive behavior supports. Further, none of the studies reviewed mentioned or reported a power analysis – which is essential "to describe the adequacy of the minimum cell size" (Gersten et al., 2005, p. 161).

SUMMARY

Clearly, the goal of creating safer, more academically productive middle and high schools is a noble and necessary focus – particularly in light of the fact that our students are living in societies that seduce our youth with violence, early sexual experimentation, and hedonistic values (Walker et al., 2004). Further, it is logical to focus our initial efforts on primary interventions that involve all students just by virtue of attending schools if the goal to prevent harm.

Yet, as this line of inquiry is developed, careful considerations must be made when designing and proposing school-wide, primary level intervention studies. In some instances, it may be that "the cost of designing a study capable of detecting small effects may simply not be interesting or worth it ... [yet,] special education research can conduct studies that are correctly powered relative to available resources, student population sizes, and goals of the research when hypothesizing moderate-to-large effects keeping the number of experimental conditions small (i.e., two rather than three groups)" (Gersten et al., 2005, p. 162). As a field, our challenge is to advance the methodological rigor as we move toward an increased emphasis on randomized trial, yet maintain an awareness of resource considerations as we continue to explore the most efficient, effective methods of designing, implementing, and evaluating school-wide, primary level interventions in secondary schools.

REFERENCES

Alspaugh, J. W. (1998). Achievement loss associated with the transition to middle school and high school. *Journal of Educational Research*, *92*, 20–25.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*, 91–97.

Carter, E., Lane, K. L., Pierson, M., & Glaeser, B. (2006). Self-determination skills and opportunities of transition-age youth with emotional disturbances and learning disabilities. *Exceptional Children*, *72*, 333–346.

Colvin, G., Kameenui, E. J., & Sugai, G. (1993). Reconceptualizing behavior management and school-wide discipline in general education. *Education and Treatment of Children*, *16*, 361–381.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., & Degimencioglu, S. M. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal*, *36*, 543–597.

DeVoe, J. F., Peter, K., Kaufman, P., Ruddy, S. A., Miller, A. K., Plany, M., Snyder, T. D., & Rand, M. R. (2003). *Indicators of school crime and safety: 2003*. Washington, DC: National Center for Educational Statistics, U.S. Departments of Education and Justice.

DuRant, R., Treiber, F., Getts, A., McCloud, K., Linder, C. W., & Woods, E. R. (1996). Comparison of two violence prevention curricula for middle school adolescents. *Journal of Adolescent Health*, *19*, 111–117.

Dwyer, K. P., Osher, D., & Warger, W. (1998). *Early warning, timely response: A guide to safe schools*. Washington, DC: U.S. Department of Education.

Fabes, R. A., Martin, C. L., Hanish, L. D., & Updegraff, K. A. (2000). Criteria for evaluating the significance of development research in the twenty-first century: Force and counterforce. *Child Development*, *71*, 212–221.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, *31*, 4–14.

Fuchs, D., & Fuchs, L. (1994). Inclusive schools movement and the radicalization of special education reform. *Exceptional Children*, *60*, 294–309.

Gainer, P. S., Webster, D. W., & Champion, H. R. (1993). A youth violence prevention program: Description and preliminary evaluation. *Archives of Surgery*, *128*, 303–308.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Gottfredson, D. C., Gottfredson, G. D., & Hybl, L. G. (1993). Managing adolescent behavior: A multiyear, multischool study. *American Educational Research Journal*, *30*, 179–215.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review*, *18*, 37–50.

Gresham, F. M., Sugai, G., Horner, R., Quinn, M., & McInerney, M. (1998). *Classroom and schoolwide practices that support students' social competence: A synthesis of research*. Washington, DC: Office of Special Education Programs.

Harnett, P. H., & Dadds, M. R. (2004). Training school personnel to implement a universal school-based prevention of depression program under real-world conditions. *Journal of School Psychology*, *42*, 343–357.

Hawkins, J. D., Catalano, R. F., Kosterman, R., Abbott, R., & Hill, K. G. (1999). Preventing adolescent health-risks behaviors by strengthening protection during childhood. *Archives of Pediatric Adolescent Medicine*, *153*, 226–234.

Hetzroni, O. E. (2003). A positive behaviour support: A preliminary evaluation of a school-wide plan for implementing AAC in a school for students with intellectual disabilities. *Journal of Intellectual and Developmental Disability*, *28*, 283–296.

Horner, R. H., & Sugai, G. (2000). School-wide behavior support: An emerging initiative. *Journal of Positive Behavior Interventions*, *2*, 231–232.

Hunter, L., Elias, M. J., & Norris, J. (2001). School-based violence prevention challenges and lessons learned from an action research project. *Journal of School Psychology*, *39*, 161–175.

Individuals with Disabilities Education Act Amendments of 1997 (1997). Pub. L. No. 105–17, Section 20, 111 Stat. 37. Washington, DC: U.S. Government Printing Office.

Individuals with Disabilities Education Improvement Act of 2004 (2004), 20 U.S.C. 1400 *et esq.* (reauthorization of Individuals with Disabilities Act 1990).

Isakson, K., & Jarvis, P. (1999). The adjustment of adolescents during the transition into high school: A short term longitudinal study. *Journal of Youth and Adolescence*, *28*, 1–26.

Kartub, D. T., Taylor-Greene, S., March, R. E., & Horner, R. H. (2000). Reducing hallway noise: A systems approach. *Journal of Positive Behavior Intervention*, *2*, 179–182.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, *1*, 427–452.

Kincaid, D., Knoster, T., Harrower, J. K., Shannon, P., & Bustamante, S. (2002). Measuring the impact of positive behavior support. *Journal of Positive Behavior Support*, *4*, 109–117.

Kern, L., & Manz, P. (2004). A look at current validity issues of school-wide behavior support. *Behavioral Disorders*, *30*, 47–59.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods* (3rd ed.). Belmont, CA: Books/Cole.

Lane, K. L., & Beebe-Frankenberger, M. (2004). *School-based interventions: The tools you need to succeed.* Boston, MA: Pierson Education, Inc.

Lane, K. L., Beebe-Frankenberger, M. E., Lambros, K. M., & Pierson, M. (2001). Designing effective interventions for children at-risk for antisocial behavior: An integrated model of components necessary for making valid inferences. *Psychology in the Schools*, *38*, 365–379.

Lane, K. L., Bocian, K. M., MacMillian, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential – but often forgotten – component of school-based interventions. *Preventing School Failure*, *48*, 36–43.

Lane, K. L., Gresham, F. M., & O'Shaughnessy, T. E. (2002). Identifying, assessing, and intervening with children with or at risk for behavior disorders: A look to the future. In: K. L. Lane, F. M. Gresham & T. E. O'Shughnessy (Eds), *Interventions for children with or at risk for emotional and behavioral disorders* (pp. 317–326). Boston: Allyn & Bacon.

Lane, K. L., Pierson, M., & Givner, C. C. (2004). Secondary teachers' views on social competence: Skills essential for success. *Journal of Special Education*, *38*, 174–186.

Lane, K. L., Umbreit, J., & Beebe-Frankenberger, M. (1999). A review of functional assessment research with students with or at-risk for emotional and behavioral disorders. *Journal of Positive Behavioral Interventions*, *1*, 101–111.

Lane, K. L., & Wehby, J. (2002). Addressing antisocial behavior in the schools: A call for action. *Academic Exchange Quarterly*, *6*, 4–9.

Lewis, T. J., & Sugai, G. (1999). Effective behavior support: A systems approach to proactive school-wide management. *Focus on Exceptional Children*, *31*, 1–24.

Loeber, R., Green, S. M., Lahey, B. B., Frick, P. J., & McBurnett, K. (2000). Findings on disruptive behavior disorders from the first decade of the developmental trend study. *Clinical Child and Family Psychology Review*, *3*, 37–59.

Lohrmann-O'Rourke, S., Knoster, T., Sabatine, K., Smith, D., Horvath, B., & Llewellyn, G. (2000). School-wide application of PBS in the Bangor area school district. *Journal of Positive Behavior Interventions*, *2*, 238–240.

Luiselli, J. K., Putnam, R. F., & Sunderland, M. (2002). Longitudinal evaluation of behavior support intervention in a public middle school. *Journal of Positive Behavior Intervention*, *4*, 182–188.

MacMillan, D., Gresham, F., & Forness, S. (1996). Full inclusion: An empirical perspective. *Behavioral Disorders*, *21*, 145–159.

Mehas, K., Boling, K., Sobieniak, J., Burke, M. D., & Hagan, S. (1998). Finding a safe haven in middle school. *Teaching Exceptional Children*, *30*, 20–23.

Metzler, C. W., Biglan, A., Rusby, J. C., & Sprague, J. R. (2001). Evaluation of a comprehensive behavior management program to improve school-wide positive behavior support. *Education and Treatment of Children*, *24*, 448–479.

Miller, D. N., George, M. P., & Fogt, J. B. (2005). Establishing and sustaining research-based practices at Centennial School: A descriptive case study of systemic change. *Psychology in the Schools*, *42*, 553–567.

Morris, R. J., Shah, K., & Morris, Y. P. (2002). Internalizing behavior disorders. In: K. L. Lane, F. M. Gresham & T. E. O'Shaughnessy (Eds), *Interventions for children with or at risk for emotional and behavioral disorders* (pp. 223–241). Boston, MA: Allyn and Bacon.

Morrison, G. M., Robertson, L., Laurie, B., & Kelly, J. (2002). Protective factors related to antisocial behavior trajectories. *Journal of Clinical Psychology*, *58*, 277–290.

Nakasato, J. (2000). Data-based decision making in Hawaii's behavior support effort. *Journal of Positive Behavior Support*, *2*, 247–251.

Nersesian, M., Todd, A. W., Lehmann, J., & Watson, J. (2000). School-wide behavior support through district-level system change. *Journal of Positive Behavior Support*, *2*, 244–247.

Netzel, D. M., & Eber, L. (2003). Shifting from reactive to proactive discipline in an urban school district: A change in focus through PBIS implementation. *Journal of Positive Behavior Interventions*, *5*, 71–79.

Pepler, D. J., Craig, W. M., Ziegler, S., & Charach, A. (1994). An evaluation of an anti-bullying intervention Toronto schools. *Canadian Journal of Community Mental Health*, *13*, 95–110.

Rosenberg, M. S., & Jackman, L. A. (2003). Development, implementation, and sustainability of comprehensive school-wide behavior management systems. *Intervention in School and Clinic*, *39*, 10–21.

Safe and Drug-Free Schools and Communities Act of 1994. (1994). Pub. L. No. 103–382, 4001–4133, 108 Stat. 3518 (codified as amended at 20 U.S.C. 7101–7143 [2000]).

Satcher, D. (2001). *Youth violence: A report of the Surgeon General*. Washington, DC: Office of the Surgeon General, U.S. Department of Health & Human Services.

Shapiro, J. P., Burgoon, J. D., Welker, C. J., & Clough, J. B. (2002). Evaluation of the peacemakers program: School-based violence prevention for students in grades four through eight. *Psychology in the Schools*, *39*, 87–100.

Skiba, R., & Peterson, R. (2003). Teaching social curriculum: School discipline as instruction. *Preventing School Failure*, *47*, 66–73.

Sprague, J., Walker, H., Golly, A., White, K., Myers, D. R., & Shannon, T. (2001). Translating research into effective practice: The effects of a universal staff and student intervention on indicators of discipline and school safety. *Education and Treatment of Children*, *24*, 495–511.

Stevens, V., De Bourdeaudhuij, I., & Van Oost, P. (2000). Bullying in Flemish schools: An evaluation of anti-bullying intervention in primary and secondary schools. *British Journal of Educational Psychology*, *70*, 195–210.

Sugai, G., & Horner, R. H. (2002a). Introduction to the special series on positive behavior supports in schools. *Journal of Emotional and Behavioral Disorders*, *10*, 130–135.

Sugai, G., & Horner, R. (2002b). The evolution of discipline practices: School-wide positive behavior supports. *Child and Family Behavior Therapy*, *24*, 23–50.

Taylor-Greene, S., Brown, D., Nelson, L., Longton, J., Gassman, T., Cohen, J., Swartz, J., Horner, R. H., Sugai, G., & Hall, S. (1997). School-wide behavioral support: Starting the year off right. *Journal of Behavioral Education*, *7*, 99–112.

Taylor-Greene, S. J., & Kartub, D. T. (2000). Durable implementation of school-wide behavior support. *Journal of Positive Behavior Support*, *2*, 233–235.

Turnbull, A., Edmonson, H., Griggs, P., Wickham, D., Sailor, W., Freeman, R., Guess, D., Lassen, S., McCart, A., Park, J., Riffel, L., Turnbull, R., & Warren, J. (2002). A blueprint for school-wide positive behavior support: Implementation of three components. *Exceptional Children*, *58*, 377–402.

U.S. Department of Justice. (2001). *Crime in the United States: 2001*. Washington, DC: Author.

Walker, H. M. (2003). *Comments on accepting the outstanding leadership award from the Midwest Symposium for leadership in behavior disorders*. Kansas City, KS: Author.

Walker, H. M., Ramsey, E., & Gresham, F. M. (2004). *Antisocial behavior in school: Evidence-based practices* (2nd ed.). Belmont, CA: Wadsworth.

Walker, H. M., & Severson, H. (2002). Developmental prevention of at-risk outcomes for vulnerable antisocial children and youth. In: K. L. Lane, F. M. Gresham & T. E. O'Shaughnessy (Eds), *Interventions for children with or at risk for emotional and behavioral disorders* (pp. 177–194). Boston, MA: Allyn and Bacon.

Warren, J. S., Edmonson, H. M., Griggs, P., Lassen, S. R., McCart, A., Turnbull, A., & Sailor, W. (2003). Urban applications of school-wide positive behavior support. *Journal of Positive Behavior Supports*, *5*, 80–91.

Warren, J., Edmonson, H., Turnbull, A., Sailor, W., Wickham, D., Griggs, P., & Beech, S. (in press). School-wide application of positive behavior supports: Implementation and preliminary evaluation of PBS in an urban middle school. *Educational Psychology Review*.

White, R., Marr, M. B., Ellis, E., Audette, B., & Algozzine, B. (2001). Effects of a model of school-wide discipline on office referrals. *The Journal of At-Risk Issues*, *7*, 4–12.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*, 203–214.

This page is left intentionally blank

EXAMINING THE EFFECTS OF PROFESSIONAL DEVELOPMENT TO IMPROVE EARLY READING INSTRUCTION: HOW STRONG IS THE CAUSAL CHAIN?

Stephanie Al Otaiba, Jeanine Clancy-Menchetti and Christopher Schatschneider

ABSTRACT

More than ever before, researchers and policymakers expect general education classroom to be the first line of defense in efforts to prevent reading difficulties. Preventing reading difficulties through evidence-based beginning reading instruction research features prominently in the 2002 No Child Left Behind legislation (NCLB; P. L. 107-110) and in the 2004 amendments to the Individuals with Disabilities Act (IDEA). The purpose of this chapter is to describe the experimental and quasi-experimental methodological approaches that have been used to examine the effects of professional development in reading on teachers' instructional practices and students' reading outcomes and to evaluate the chain of causal linkage in the more recent studies. The first section of the chapter provides a brief history of relevant research. The second section summarizes findings of

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 201–236

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19008-4

the National Reading Panel (NRP, 2000) Report and those of a recent review of the literature (Clancy-Menchetti & Al Otaiba, 2006). The final section synthesizes what we have learned from the research.

The ability to read proficiently not only improves learning in a variety of content areas throughout children's school careers, but also increases employment opportunities and overall quality of life. Given increasing societal demands for a highly literate workforce (Adams, 1990; Puma et al., 1997), there is wide agreement that far too many children and adults can not read proficiently enough to be successful (National Assessment of Educational Progress, 2003; Snow, Burns, & Griffin, 1998). Preventing reading difficulties is vital because reading difficulties become increasingly intractable after third grade and there is considerable evidence that poor readers rarely catch up to good readers over the elementary years (Allington, 2002; Good, Simmons, & Kame'enui, 2001; Juel, 1988; Kennedy, Birman, & Demaline, 1986; Lyon, 1985; National Center for Educational Statistics, 2002; Spira, Bracken, & Fischel, 2005; Stanovich, 1986). Moreover, reading difficulties remain the most common reason for referral for special education among the over 2.8 million children with specific learning disabilities (e.g., Mastropieri, Leinart, & Scruggs, 1999; President's Commission on Excellence in Special Education, 2002).

Converging findings reported in several influential reviews of empirical reading research (Adams, 1990; National Reading Panel (NRP), 2000; Snow et al., 1998; Snow, 2002) agree that the best way to prevent reading difficulties is to ensure that all children receive explicit and systematic beginning reading instruction, that includes five core components: phonological awareness, phonics, fluency, vocabulary, and comprehension. Researchers also emphasize that access to evidence-based instruction is especially critical for children who enter school with limited literacy experience, who may be more sensitive to the quality of instruction than children who enter school with richer literacy experience (Foorman et al., 1997; Goldhaber & Anthony, 2004).

More than ever before, researchers and policymakers expect general education classroom to be the first line of defense in efforts to prevent reading difficulties. Preventing reading difficulties through evidence-based beginning reading instruction research features prominently in the No Child Left Behind legislation (NCLB, 2001; P. L. 107–110) and in the 2004 amendments to the Individuals with Disabilities Act (IDEA). Under both legislative mandates, general education classroom teachers are expected to provide children with evidence-based classroom reading instruction, monitor children to identify those who do not make adequate progress, and to provide these children with more and intensive scientifically based intervention. This multi-tier process of increasingly intensive levels of education is termed responsiveness to intervention (RTI). Ideally, the purpose of RTI is to reduce the numbers of students needing special education services (see for example Fuchs & Fuchs, 1998; Vaughn, Linan-Thompson, & Hickman, 2003).

Yet, many general educators (and special educators) have reported feeling unprepared to teach reading, especially to children attending high-poverty schools (Moats & Lyon, 1996; Lewis et al., 1999). Certainly as a group, educators are highly skilled in reading and spelling, but their own skill does not ensure that they feel equipped to use this knowledge to assist struggling readers. For example, McCutchen et al. (2002) found no relationship between teachers' knowledge of children's literature and their ability to conduct comprehension or writing activities to struggling beginning readers.

Given the vital role of classroom teachers in early reading instruction and intervention within the RTI framework, we agree with Sweet and Snow's (2003) recommendation that teachers, and particularly those teachers working in high-poverty schools, "need guidance about how to combine and prioritize various instructional approaches in the classroom and in particular about how to teach comprehension while attending to the often poor word-reading skills their students bring … ." (pp. 47–48) A requirement under the NCLB (P. L. 107-110, 2002) is that teachers receive "high-quality" professional development. Over $3 billion in federal funding has been allocated to this effort. The recent 2002 report, *Specific Learning Disabilities: Finding Common Ground* (2002) stated that

> At the core of a high-quality education is effective delivery of appropriate research-based interventions by teachers and other professionals, and on-going monitoring and assessment coordinated by interdisciplinary teams … . (P)ositive results and improvement will not occur unless teachers and other professionals in the system have the knowledge, skills, and administrative support to implement these new measures within a collaborative system that brings regular and special educators, related services personnel, and administrators together. (p. 4)

Nearly 20 years ago, Showers, Joyce, and Bennett (1987) conducted a synthesis of the staff development literature and proposed a "package" of guiding principles for high-quality professional development. These principles for improving teachers' instructional practices have been widely cited and are still widely used (see for example, Brady & Moats, 1997; Cochran-Smith & Lyttle, 1999; *Every child Reading: A Professional Guide*, Learning First Alliance, 2000; Office of Educational Research and Improvement, 1997). These principles include (a) training in the theoretical framework for the professional development, (b) demonstrating or modeling the practices,

(c) providing opportunities for practice and guidance, and (d) coaching or mentoring. Showers et al. also emphasized that professional development should be ongoing and embedded in teacher's daily routines.

However, given the tremendous cost of professional development and given the high stakes for children and teachers, we are concerned that there is surprisingly little rigorous research that provides causal information linking these principles of professional development with improved outcomes for general education teachers and for students, or more specifically, for those students considered at-risk for future reading difficulties (Whitehurst, 2002). The purpose of this chapter is to (a) describe the experimental and quasi-experimental methodological approaches that have been used to examine the effects of professional development in reading on teachers' instructional practices and students' reading outcomes and (b) to evaluate the chain of causal linkage in the more recent studies. The first section of the chapter provides a brief history of relevant research. The second section summarizes findings of the NRP (2000) report and those of a recent review of the literature (Clancy-Menchetti & Al Otaiba, 2006). The final section synthesizes what we have learned from the research. It provides an exemplary study that links professional development to (a) improved teacher knowledge, attitude, and instruction and (b) improved student outcomes on multiple measures of the components taught during the professional development process. The section concludes with a discussion of implications for future research and practice.

BRIEF HISTORY OF THE EVOLUTION OF RESEARCH METHODOLOGY USED TO EXPLORE THE RELATIONSHIP BETWEEN INSTRUCTION AND STUDENT LEARNING

Many people in the United States felt we had lost the race to outerspace when the Soviet Union launched Sputnik in 1957. The inherent vulnerability felt as a result of this single event, sparked a wave of public concern about the quality of teaching, particularly for children living in poverty. The large-scale Coleman Report (Coleman et al., 1966) was commissioned by Congress to examine the relationships among teachers, resources, and student achievement within the context of racially segregated and integrated schools. Coleman et al. (1966) found that the achievement of minority children who attended segregated schools was behind white children and this gap steadily increased over their school careers. By contrast, minority children had higher achievement when they attended integrated and middle class schools. This seminal report established a baseline for the notion that teachers play a critical role in student achievement and spawned a line of research focusing on identifying elements of exceptional schools and effective teachers that could "beat the odds". The report also led to increased interest in studying of the impact of teacher quality and teacher education on student outcomes.

Throughout the 1960s and 1970s, process-product research, which focused on identifying the teaching behaviors that were correlated to student achievement, was prevalent (Brophy, 1973; Rosenshine, 1979). Researchers established that several teaching behaviors were correlated or associated with high student outcomes including: clarity, enthusiasm, high expectations, task orientation, small group instruction, directive teaching (modeling, guided practice, followed by independent practice), immediate feedback, student engagement, and opportunity to learn or time spent in academic skill instruction (Brophy & Good, 1986; Hoffman, 1991; Rosenshine & Furst, 1973).

As early as 1973, Rosenshine and Furst called for a paradigm shift from correlational research to more rigorous experimental research that would provide causal information about how to improve beginning reading instruction during professional development efforts. Shortly thereafter, Dunkin and Biddle (1974) specified four categories of variables they hypothesized would be useful in such research. The first category pertains to the independent variable, the *process* of professional development or the *process* of instruction (e.g., instructional strategies, small group instruction). The second category, *presage* variables, includes the knowledge and characteristics of teachers (e.g., knowledge, cognitive ability, educational background, and personality). The third category describes *product,* defined as pupil achievement (e.g., pupil learning measured as growth or gain scores) and teacher improvement (e.g., subject matter knowledge, change in beliefs or attitudes). The fourth category of variables related to *context* variables, refers to the school conditions over which the teacher has no control (e.g., child characteristics including socioeconomic status, curriculum, class size, school leadership, organization of school routines and schedules).

Over the past three decades, researchers began to explore more thoroughly the importance of Dunkin and Biddle's (1974) four categories of variables within the broader field of education. The importance of *context* variables such as efficient school instructional leadership and use of data-based decision-making have been well documented in the school reform literature (e.g., Weber, 1971; Wilder, 1977; Venezky & Winfield, 1979). Schereens and Bosker (1997) pitted school *context* variables with classroom variables (*presage* and *process*) and reported that the classroom variables account for 15–20% of the unique variance in overall student achievement. Darling-Hammond (2000a) has documented that *presage* variables, particularly those related to teacher characteristics (e.g., preservice teaching preparation, certification, and professional development) explain the largest variance in student reading achievement on the National Assessment of Educational Progress, even after controlling for *context* variable such as student poverty and language background. Similarly, when Rivers and Sanders (2002) recently examined factors related to student achievement, they concluded that instructional quality (*process* variables) contributes more than ethnic and socioeconomic characteristics, previous achievement, or class size.

A large body of qualitative research has also described aspects of the *process* of effective teaching in hundreds of high-achieving elementary-level classrooms (Knapp, 1995; Pressley et al., 2001; Taylor, Pearson, Clark, & Walpole, 1999; Wharton-McDonald, Pressley & Hampston, 1998). The most effective teachers (those whose children had the highest gains in reading) provided: "more small group instruction, communicated more with parents, had children engage in more independent reading, provided more coaching during reading as a way to help children apply phonics knowledge, and asked more higher-level questions" than less-effective teachers (Taylor, Pressley, & Pearson, 2002, p. 365). Taylor et al. also emphasized that the most effective teachers worked in the *context* of effective schools. These effective schools shared certain school-wide context variables that supported teaching and learning: "focused on improved student learning, strong school leadership, strong teacher collaboration, consistent use of data on student performance, focus on professional development and innovation, strong links to parents" (p. 369).

Shulman (1986) conducted a review that examined the role of *presage* variables and found that teachers' beliefs and their content knowledge, or subject matter knowledge, was by and large the "missing paradigm" (p. 6). Since his review, researchers have begun to examine more systematically what teachers need to know in order to teach reading well (Grossman, 1990, 1991; Moats, 1994; Moats & Lyon, 1996; McCutchen & Berninger, 1999). More recently, as researchers have realized that teacher knowledge of scientifically based research is limited but of critical importance in bridging the research to practice gap, they have begun to focus on teacher *product* variables, including improving teachers' knowledge and changing their beliefs which has led to the development of measures of content knowledge for teaching reading (Moats, 2000; Mather, Bos, & Babur, 2001; Phelps & Schilling, in press). Thus, each of these four categories of variables has the potential to impact the linkage among professional development and teacher and student outcomes. It is important to analyze the degree to which researchers have measured and manipulated process, presage, product, and context. To aid in our investigation of these variables as they relate to professional development, we have examined the literature on professional development starting with the NRP report on teacher education and reading instruction.

FINDINGS FROM THE NATIONAL READING PANEL AND A MORE RECENT REVIEW OF THE LITERATURE

In this section of the chapter, first we describe the criteria used for selecting studies in both reviews. Then, we briefly summarize their findings.

Criteria for Selecting Studies and a Brief Summary of Findings

Selection criteria used in the reviews. The NRP (2000) followed strict criteria for selecting studies in their review of the literature, which we also followed in conducting our more recent review (Clancy-Menchetti & Al Otaiba, 2006). These criteria included: (a) described professional development related to beginning reading instruction; (b) measured both teacher and student outcomes; (c) were published in peer-reviewed journals; (d) used an experimental or quasi-experimental research design; and (e) took place in the United States. Additionally, because our interest related to professional development designed to help teachers learn how to prevent reading difficulties through effective early reading instruction, we selected only studies that focused on grades pre-k through primary grade. Although the NRP located four studies related to teacher preparation of teachers for comprehension strategy instruction and additional 21 studies related to teacher education and reading instruction, only 13 of these focused on inservice teacher training and provided both teacher and student outcome measures. We located an additional nine studies.

We were surprised to find so few relevant articles given our extensive review, which included the following steps: (1) conducting an electronic search of the Educational Research Information Center (ERIC) and PsychINFO databases, (2) hand searching the following journals from 1999 to the present: *American Education Research Journal*, *Annals of Dyslexia, Childhood Education*, *Early Childhood Research Quarterly*, *Educational Leadership*, *Elementary School Journal*, *Journal of Educational Psychology*, *Journal of Educational Research*, *Innovative Higher Education*, *Journal of Learning Disabilities*, *Learning Disabilities Research & Practice*, *Literacy Research*, *Journal of Research and Development in Education*, *Reading Improvement*, *Reading Teacher*, *Reading Research Quarterly*, *Review of Educational Research*, *and Teaching and Teacher Education*, Teacher Education and Special Education, and (3) examining handbooks and relevant texts for supplemental resources and conducting an extensive ancestral search (foot-note chasing).

Brief summary of findings from both reviews. A detailed description of the findings of these two separate reviews is beyond the page limitations and the scope of this chapter, so what follows is a brief synthesis and a helpful summary in Table 1, which includes the type of experimental design (i.e., experimental or quasi-experimental), whether there was a control group or not, if pre- and posttests were administered, and a short description of findings related to teacher and student outcomes.

When the NRP's Teacher Education and Reading Instruction and Teacher Preparation Subgroup (2000) examined the relevant literature, they found that teachers learned to implement what they were taught and that their students generally improved their reading skills (we refer the reader to the NRP's Teacher Education and Reading Instruction Subgroup's Report (Table 4: Inservice studies with teacher and student outcome measures, pp. 5–32)). The NRP cautioned, however, that due to gaps and methodological concerns about the existing research base, they were unable to conduct a meta-analysis and, they were unable to directly or causally link professional development to teacher change and to student improvement.

Indeed, as we studied the NRP (2000) report and we thoroughly read those investigations reviewed by the Panel, we found that only slightly more than half (7 of the 13 studies) demonstrated that treatment teachers performance was statistically significantly different than control teachers. Only one team of investigators reported being unable to show teachers whether treatment teachers could implement training; but that may not be surprising given that professional development was provided only in a "self-study" format (Coladarci & Gage, 1984). An additional study (Stallings & Krasavage, 1986) found that teachers' observed implementation actually decreased, but this decrease occurred during a "sustainability" year when there

Table 1. Teacher and Student Findings from the 22 Studies (13 National Reading Panel Studies and 9 More Recent Studies).

Author/s & Date	Research Design and Method of Analysis	Professional Development Effectiveness	
		Teacher findings	Student findings
	13 Studies Pre-National Reading Panel Report		
Anderson, Evertson, and Brophy (1979)	Quasi-experimental study[a] with control group. Posttest measures only for students ANOVA	Treatment teachers used more direct instruction strategies than control teachers[b]	Treatment students had significantly higher adjusted achievement scores however, school effects cannot be ruled out
Baker (1977)	Quasi-experimental study with control group for students only. Pre- and post-measures for students only	Teacher report change in self awareness and in questioning techniques	Treatment students gain scores significantly higher on Gilmore Oral Reading[c] subtests of accuracy and comprehension and Schonell Graded Word Reading List[d]
Book, Duffy, Roehler, Meloth, and Vavrus (1985)	Experimental study with control group. Posttest measures only	Treatment teachers were observed to be significantly more explicit in their explanations and more explicit over time	Treatment students scored significantly higher on awareness of strategies
Brown, Pressley, Van Meter, and Schuder (1996)	Quasi-experimental study with control group. Pre- and post-measures for students only. Used gain scores to report outcomes	Treatment teachers observed to have more prominent discussion of strategies	Treatment students scored significantly higher on comprehension and word skills subtests of Stanford Achievement Test,[e] and were significantly more interpretive in story retell Treatment students reported more awareness of comprehension and word-level strategies
Coladarci and Gage (1984)	Experimental study with control group. Pre- and post-measures of teachers and students ANCOVA	Treatment implementation was poor, teachers did not conform to training recommendations	No significant difference in end-of-year student achievement between groups

Table 1. *(Continued)*

Author/s & Date	Research Design and Method of Analysis	Professional Development Effectiveness	
		Teacher findings	Student findings
Conley (1983)	Experimental study with control group. Pre- and post-measures of students only ANOVA	Qualitative description reports teachers benefited	Treatment students had significant comprehension gains on Gates–MacGinitie Reading Test[f] (level E) showing twice as much growth as control students
Duffy et al. (1986)	Experimental study with control group. Pre- and post-measures of teachers and students ANOVA and ANCOVA	Treatment teachers significantly higher in explicit strategy instruction	Treatment students demonstrated significantly more awareness of reading strategies based on interview ratings No achievement gains in comprehension as measured by Gates-MacGinitie Reading Test[f]
Duffy et al. (1987)	Experimental study with control group. Pre- and post-measures of teachers and students ANOVA, ANCOVA, MANOVA, and MANCOVA	Treatment teachers significantly higher in explicit strategy instruction	Treatment students scored significantly higher on word meaning and word recognition subtests of Stanford Achievement Test[e] and on Michigan Educational Assessment Program[g] and in lesson interviews (situational and procedural knowledge) and concept interviews of a supplemental achievement measure (researcher designed)
Miller and Ellsworth (1985)	Quasi-experimental and experimental (with sub group) with control group. Pre- and post-measures with teachers only. Reported gain scores ANOVA	Significant post test difference on knowledge test favoring treatment teachers; three out of 28 items on attitude survey showed significant differences favoring treatment teachers Treatment teachers had higher implementation levels	Posthoc analysis indicated significant differences favoring treatment students on the reading comprehension subtest of the California Achievement Test[h]

Stallings, Robbins, Presbrey, and Scott (1986)	Quasi-experimental with control group. Pre- and post-measures for teachers and students. Reported gain scores and *t* tests results	Treatment teachers significantly improved instructional skills based on observation instrument	Students significantly improved engaged rates in reading Lmited English speaking students were significantly more engaged in reading than English-speaking students
Stallings and Krasavage (1986)	Quasi-experimental with control group. Pre- and post-measures for teachers and students. Reported gain scores and *t*-tests results ANCOVA	Ratings on observation instruments dropped for 7 of 10 teachers in both subjects	Engaged rates in both subjects dropped significantly Greater gains for control students on standardized tests; limited English speaking students gained more than English-speaking students
Streeter (1986)	Experimental study with control group. Pre- and post-measures of teachers and students. Reported gain scores and *t* tests results ANCOVA	Treatment teachers displayed significantly higher levels of enthusiasm than control teachers	Treatment students scores on expressed reading difficulty decreased significantly
Talmage, Pascarella, and Ford (1984)	Quasi-experimental with control group. Pre- and post-measures of teachers and students ANOVA and ANCOVA	Treatment teachers had significantly more positive attitudes towards cooperative learning strategies and observed cooperative practices	Some positive effects for reading but not language arts when scores pooled
	Nine Studies Post-National Reading Panel Report		
Baker and Smith (1999)	Quasi-experimental study with control group. Pre- and post-measures for teachers and students ANOVA	Reported positive changes in teacher behavior	Treatment students significantly higher on segmentation At-risk students made significant gains in letter names and sounds but did not catch up to their grade level peers; during sustainability year at-risk students surpassed peers on phonemic segmentation
Bos et al. (1999)	Quasi-experimental study with control group. Pre- and post-measures of teachers and students Repeated measures ANOVA	Treatment teachers showed significant growth on teacher knowledge survey Teachers reported professional development had powerful impact on their teaching practices and were	Students from professional development teachers had significantly greater gains in all grade levels. Kindergarten – sound identification, spelling of nonsense

Table 1. *(Continued)*

Author/s & Date	Research Design and Method of Analysis	Professional Development Effectiveness	
		Teacher findings	Student findings
		observed to have high levels of fidelity	words, spelling of real words; 1st grade – spelling of nonsense and real words; 2nd grade – reading and spelling
Dickinson and McCabe (2001)	Quasi-experimental study with control group. Pre- and post-measures of teachers and students HLM	Treatment teachers significantly higher on environment and tools including presence and use of books and the overall support for literacy	Treatment students scores on receptive vocabulary significantly higher than control
Dickson and Bursuck (1999)	Quasi-experimental study without control group. Pre- and post-measures of teachers and students MANOVA and MANCOVA	Teachers felt good about the Early Intervention Program	High risk students made growth but they did not catch up to peers
Greenwood, Tapia, Abbott, and Walton (2003)	Longitudinal study without control group. Pre- and post-measures of teachers and students HLM	Results based on observations – teachers implemented with fidelity, teachers arranged reading to occur more often in small groups and one-one instead of whole class	HLM all students showed growth on curriculum based measures of oral reading fluency and their rates increased over the three years
Jackson, Paratore, Chard, and Garnick (1999)	Quasi-experimental study without control group. Pre- and post-measures of teachers and students Student measures compared to benchmarks	Teacher implemented with fidelity but weak on instructional pacing, segmentation, and guided writing	Too few students to analyze Students made gains but did not catch up to grade level peers; Teacher opinion: intervention helped
McCutchen and Berninger (1999)	Quasi-experimental study with matched control group of teachers. Pre- and post-measures of teachers and students Repeated measures ANOVA HLM	Treatment teachers linguistic knowledge significantly higher; treatment teachers spent significantly more time on instruction in phonological awareness in kindergarten and more time on comprehension in 1st grade	Significant growth across grades for treatment students: kindergarten – phonological awareness, orthographic fluency, and word reading. 1st graders – phonological awareness, word reading, spelling, and composition fluency. 2nd graders – composition fluency

McCutchen et al. (2002)	Quasi-experimental study with control group. Pre- and post-measures of teachers and students Repeated measures ANOVA HLM	Same as above	Same as above
O'Connor (1999)	Quasi-experimental study with control group. Pre- and post-measures of teachers and students ANOVA and MANOVA	Teachers receiving PD[i] performed more phonological awareness activities than control teachers; number of strategies implemented did not differ by level of PD[i] intensity however more intense PD[i] teachers had higher levels of implementation	Students from PD[i] teachers scored significantly higher in all measures
Descriptive: comparing students by level of PD[i] intensity; students from more intense PD[i] teachers scored higher in letter naming, word identification, and spelling			

Abbreviations: ANOVA, analysis of Variance; ANCOVA, analysis of covaviance; MANOVA, Multianalysis of variance; MANCOVA, Multianalysis of Covariance.

[a] under the column Type, Q = quasi-experimental; E = experimental study; and L = longitudinal, sequential cohort design.

[b] Ss is students and Ts is teachers.

[c] Gilmore Oral Reading (Gilmore & Gilmore, 1968).

[d] Schonell Graded Word Reading List (Schonell & Schonell, 1960).

[e] Stanford Achievement Test (Gardner, 1982).

[f] Gates-MacGinitie Test (MacGinitie, MacGinitie, Maria, & Dreyer, 2000).

[g] Michigan Educational Assessment Program (Michigan Department of Education, 1994).

[h] California Achievement Test (CTB/Macmillan /McGraw Hill, 1992).

[i] Professional Development.

was less support for implementation. With regard to student improvement, only three investigations (including the two just-mentioned) did not find statistically significant improvement for students (Coldarci & Gage, 1984; Stallings & Krasavage, 1986; Talmage, Pascarella, & Ford, 1984).

Ironically, the existing database related to professional development was not rigorous enough to allow the NRP to make specific recommendations or to describe a set of evidence-based professional development practices to help teachers learn how to train their students in the very components the NRP found to have positive and significant effects on reading skills for at-risk children. In other words, the paucity of research related to teacher professional development contrasted sharply from the robust findings and determinations of the other subgroups regarding the positive effects of training (most frequently conducted by researchers and research staff rather than teachers) for at-risk children in the components of phonological awareness, phonics, vocabulary, fluency, and comprehension. The NRP concluded that teachers "need extensive support (both time and money) on a continuing basis" (pp. 5–13) and called for further rigorous research that would allow causal conclusions to be made in the area of teacher education.

In our review of nine more recent studies (Clancy-Menchetti & Al Otaiba, 2006), we were encouraged to find that the professional development process began to include training that reflect other subgroup findings regarding the importance of explicit and systematic instruction in the five components of phonological awareness, phonics, vocabulary, fluency, and comprehension. Such work is important to bridge the research to practice gap. Our findings related to teachers and students concurred with the NRP (2000) – teachers generally learned to implement what they were taught and professional development was generally associated with positive improved student outcomes. Nevertheless, we refer the reader to Table 1 and caution that only four of the nine studies we reviewed demonstrated that treatment teachers performed statistically significantly better than control teachers. The remaining studies provided qualitative descriptions of observations that suggest teachers were in fact implementing the program. Encouragingly, only one study noted that a single teacher's implementation as "weak", and only in one narrow aspect of implementation (Dickson & Bursuck, 1999). Five of the nine teams of investigators reported statistically significant findings on a range of reading and spelling measures that favored students whose teachers received professional development. The remainder reported growth in the desired direction, or that there were too few students to analyze results statistically.

Need for closer examination of causal linkage. In summary, it appears that professional development holds promise, but in order to determine whether professional development has a causal effect on teacher behavior and student outcomes, researchers must establish clear linkage that leads directly from the independent treatment variables (professional development process) to the dependent variables (teacher product, and from there to student product). There are numerous texts that the reader is no doubt familiar with that explain more thoroughly concepts related to research design and internal and external validity (cf. Shadish, Cook, & Campbell, 2002) and provide recommendations on how to conduct rigorous educational research (e.g., Cochran- Smith & Zeichner, 2005; NRP, 2000; Shavelson & Towne, 2002; Vellutino & Schatschneider, 2004). For the purposes of this chapter, an important distinction is the difference between experimental and quasi-experimental studies. Generally speaking, experimental designs, involving random assignment of participants to condition, protect against threats to internal validity and can strengthen researchers' causal claims. However, as many of us have found in our own work, experiments are not always feasible in educational settings (e.g., Shadish et al., 2002; Vellutino & Schatschneider, 2004). Indeed, only 6 of the 22 studies in the existing research base utilized randomized treatment control experimental designs.

The remaining 16 studies, including all 9 of the more recent studies, used quasi-experimental approaches. In quasi-experiments, participants are not assigned randomly to condition; therefore, researchers must take steps to attempt to distribute "potential confounds equally across all groups" (Vellutino & Schatschneider, 2004, p. 129). The ability of quasi-experimental designs to accurately estimate causal effects are related to the design features implemented in the study (Heinesman & Shadish, 1996). Some prominent features of quasi-experimental designs that increase the accuracy of effect size estimation include (a) pretest equivalence at baseline, (b) minimize the ability of participants to self-select into either the treatment or control group, (c) insure that your measurement instruments are highly reliable and valid, and (d) minimize extraneous factors that could confound the causal relationship between treatment and effect (i.e., history, selection bias, instrumentation, attrition). Thus, to evaluate the strength of the potentially causal relationship between professional development and student outcome we asked ourselves two questions.

1. How thoroughly have treatment, moderator, and dependent variables been evaluated?
2. How strong are the links of the causal chain of evidence between treatment and teacher and student improvement within the more recent studies?

EXAMINING THE PROCESS, PRESAGE, PRODUCT, AND CONTEXT VARIABLES AND METHODS OF DATA ANALYSIS: A COMPARISON OF NRP AND MORE RECENTLY REVIEWED STUDIES

In this section, we analyze the degree to which researchers have measured and manipulated process, presage, product, and context. Specifically, these reviewing criteria are used to evaluate the causal relationship between professional development and teacher and student outcomes. In order to compare the professional development models themselves, the major components of each model have been provided in Table 2. This table also shows the linkage between the instructional topics and components taught during professional development and significant teacher and student gains on these components specifically for the nine studies we reviewed. For ease of reference, the table repeats some information seen in Table 1. First, along with the authors the table provides a brief description of the research design and method of analysis used in the study. Second, it outlines the major components in the professional development model. Third, it lists which scientifically based components were presented during the professional development process. Fourth, it details teacher measures (knowledge and observed instruction) and products, and finally it describes student measures and products. We also refer the reader once again to the analogous Table 4 prepared by the NRP's Teacher Education and Reading Instruction Subgroup's Report (2000, pp. 5–32).

Process (Treatment Variables)

Focus of professional development process. Recall that by "process" Dunkin and Biddle referred to instructional strategies and routines. With regard to process, we examined whether the professional development trained teachers to provide children with instruction in the five core components of scientifically based reading research. No studies reviewed by the NRP (2000) trained teachers in phonemic awareness, vocabulary, or reading fluency. However, the NRP did find that six investigations addressed comprehension or comprehension strategy instruction (Baker, 1977; Brown, Pressley, Van Meter, & Schuder, 1996; Conley, 1983; Duffy et al., 1986; Duffy et al., 1987; Miller & Ellsworth, 1985); in three of these six, teachers were also taught to build students' word recognition (Baker, 1977; Brown et al, 1996; Duffy et al., 1987).

Table 2. Linkage Among Process, Teacher Product, and Student Product.

Authors, Research Design and Method of Analysis	Professional Development Process	Focus of Professional Development	Teacher Measures	Teacher Findings	Student Measures	Student Findings
Baker and Smith (1999) Quasi-experimental with control group. Pre and post measures for teachers and students. ANOVA	1. Initial planning meetings 2. Field test new strategies in the classroom 3. Teacher requested mentoring available 4. Researcher provided feedback and reflections during weekly meetings 5. Occasional group meetings to reflect on changes made 6. Principal involved	Phonemic awareness Alphabetic understanding	Classroom observations (14) with field notes with summaries and reflections; formal and informal interviews, and fidelity checks Teachers given feedback in the form of written summaries after observations Inference level: medium	Positive changes in teacher behavior Interrater agreement (90–98%) Level of fidelity not reported	*Yopp–Singer*[a] *Test of Phoneme Segmentation Fluency (PSF)* *Dynamic Indicators of Basic Early Literacy Skills (DIBELS)*:[b] Initial Sound Fluency (ISF) and Phoneme Segmentation Fluency (PSF) *Concepts about Print*[c] measured alphabetic understanding	Statistically significant growth in favor of treatment students on PSF
Bos et al. (1999) Quasi-experimental with control group. Pre and post measures for teachers and students. Repeated measures ANOVA	1. Initial intensive training course (2 1/2 weeks) with demonstrations 2. Embedded mentors collaborated monthly with individual teachers 3. Monthly follow-up sessions (1hour) for new information and discussion	Development of early reading and spelling and the structure of spoken language Phonological awareness Assessments to identify early difficulties in reading and spelling Strategies for teaching reading and spelling	*Structure of Language-Knowledge Assessment*[d] *The Teacher Attitudes of Early Reading and Spelling*[e] Reflective journals, observations with field notes, interviews and course evaluations Inference level: low	Treatment teachers showed significant growth on *Structure of Language-Knowledge Assessment*[d] Teachers reported professional development had powerful impact on their teaching practices and were observed to have high levels of fidelity	Informal test of letter – sound knowledge, WJ III[f] subtests: spelling, spelling of sounds, reading fluency	Statistically significant growth in favor of treatment students KG – letter-sound knowledge Spelling of sounds; 1st grade – spelling Spelling of sounds; 2nd grade – spelling Reading fluency
Dickinson and McCabe (2001)	1. Initial intensive training course (4 credits) taken by	Literacy and language development	*Early Language And Literacy Classroom*	Treatment teachers rated significantly	1. *Peabody Picture Vocabulary Test-*	Statistically significant growth

Table 2. *(Continued)*

Authors, Research Design and Method of Analysis	Professional Development Process	Focus of Professional Development	Teacher Measures	Teacher Findings	Student Measures	Student Findings
Quasi-experimental with control group. Pre and post measures for teachers and students. HLM	teachers and supervisors – two separate three day institutes	methods: using books more effectively, supporting writing, fostering language development, helping parents support literacy development	*Observation (ELLCO)*[g] Inference level: low	higher on *ELLCO*[g] subscales *Environment* and *Tools* specifically connected to language and literacy Interrater agreement (>90%)	*Revised (PPVT-R)* [h] 2. WRAT;[i] CAT[j] (reading comp) 3. Profile of early literacy development[k]	in favor of treatment students *on* PPVT-R:[h] receptive vocabulary Results sustained two years after intervention
Dickson and Bursuck (1999) Quasi-experimental with control group. Pre and post measures for teachers and students. MANOVA and MANCOVA	1. Pre meeting with teaches and principal to determine immediate needs 2. Initial training to learn specific literacy skills 3. Coaching and feedback during second half of year from research staff (twice month)	*Phonemic awareness* Evidence-based programs modified for specific needs: *Open Court*[l] *Phonological Awareness Training for Reading,*[m] *Spelling Through Phonics,*[n] *Wilson Reading System*[o]	Monthly observations and teacher interviews Inference level: high	Teachers "felt good" about the early intervention program	*Test of Phonological Awareness* (TOPA);[p] *DIBELS*[b] subtests: rapid letter naming fluency and PSF *WRMT-R*[q] subtest: word attack Informal Measures: letter-sound correspondence[r] and invented spelling[s]	MANOVA to test differences between program effectiveness High risk students made growth but they did not catch up to peers
Greenwood et al. (2003) Longitudinal sequential cohort design. Cohort 1 = 3 years 2 = 2 years 3 = 1 year Pre- and post-measures of teachers and students. HLM	1. Initial formal training session 2. Classroom demonstrations 3. Feedback on quality of implementation 4. Teacher option to request additional in class consultations 5. Second formal training session in Spring	Evidence-based literacy programs with a focus on phonemic awareness, progress monitoring	*Code for Instructional Structure and Student Academic Response: Main Stream* (MS:CISSAR)[t] Inference level: low	Teachers implemented 13 different evidence-based strategies Implementation stronger and more intensive during first two years of project Teachers implemented with fidelity (> 85%)	Curriculum based measures on ORF and observation of students' behavior	HLM to test cohort differences in growth parameters –NS All cohorts showed "substantial growth" on ORF

				Teachers arranged reading to occur more often in small groups and one-on-one instead of whole class		
Jackson, Paratore, Chard, and Garnick (1999) Quasi-experimental without control group. Pre and post measures for teachers and students. Student measures compared to benchmarks.	1. Initial formal training session with demonstrations 2. Weekly classroom visits with feedback 3. Follow-up session in Spring	Emergent and beginning reading, *Early Intervention Project* consisted of repeated reading of book of the week, guided writing, phonological awareness and phonics	Detailed field notes with researcher-made *Lesson Observation Form* with fidelity checks Inference level: low	Teacher implemented “with fidelity” but weak on instructional pacing, segmentation, and guided writing	*DIBELS:*[b] PSF and NWF ORF – running records on familiar and unfamiliar text; a district sponsored literacy performance assessment	Too few students to analyze – mean scores reported Descriptive: students made gains but did not catch up to grade level peers; teacher opinion: intervention helped
McCutchen et al. (2002) McCutchen & Berninger (1999) Quasi-experimental with control group. Pre and post measures for teachers and students Repeated measures ANOVA and HLM	1. Initial intensive training course (two week) 2. Embedded consultations in teachers’ classrooms 3. Follow-up sessions (3) 4. Researchers shared notebook of literacy activities 5. Teachers shared detailed lesson plans	PA, Phonics, Fluency, Vocabulary, Comprehension, motivation	*Informal Survey of Linguistic Knowledge*[u] Observations Inference level: medium	Treatment teachers showed significantly growth on *Informal Survey of Linguistic Knowledge.*[u] Treatment teachers spent significantly more time on instruction in PA in kindergarten and more time on comprehension in 1st grade	Tests of PA; orthographic fluency; ORF; comprehension; spelling; composition fluency	Statistically significant growth in favor of treatment studentsKG – PA, orthographic fluency, and ORF 1st graders – PA, word reading, spelling, and composition fluency 2nd graders – composition fluency
O’Connor (1999) Quasi-experimental with control group. Pre and post measures for teachers and students ANOVA and MANOVA	Intensive model 1. Initial intensive course (two weeks) 2. Follow-up sessions (12) every three weeks (after school) include discussions and modeling and practicing new skills	PA, alphabetic principle, print awareness	Weekly observations for treatment teachers and monthly for control teachers Inference level: high	Treatment teachers performed more PA activities than control teachers Comparing by level of PD[v] intensity: number of strategies	*PPVT-R;*[h] *WJII*[f] *subtest: Word Identification*; tests of short term memory; phonological manipulation; letter knowledge rhyme production;	Statistically significant growth in favor of treatment students in all measures Comparing students by level of PD[v] intensity: students

Table 2. (*Continued*)

Authors, Research Design and Method of Analysis	Professional Development Process	Focus of Professional Development	Teacher Measures	Teacher Findings	Student Measures	Student Findings
	3. Weekly embedded mentoringTraditional mode 1. Initial intensive course (two weeks) 2. Follow-up sessions (3) throughout year 3. Observed twice during year			implemented did not differ by level of PD[v] intensity however more intense PD[v] teachers had higher levels of implementation	rapid letter naming, segmenting and blending	from more intense PD[v] teachers scored higher in letter naming, word identification, and spelling

[a] Yopp–Singer Test of Phonemic Segmentation (Yopp, 1995) the primary measure to evaluate student outcomes.
[b] Dynamic Indicators of Basic Early Literacy Skills (Kaminski & Good, 1996).
[c] Concepts About Print (Clay, 1985).
[d] Structure of Language Knowledge Assessment (Bos et al., 1999).
[e] Teacher Attitudes of Early Reading and Spelling (Deford, 1985).
[f] Woodcock–Johnson Tests of Achievement III (Woodcock, McGrew, & Mather, 2001).
[g] Early Language and Literacy Classroom Observation Instrument (Smith & Dickinson, 2002).
[h] Peabody Picture Vocabulary Test-Revised (Dunn & Dunn, 1981).
[i] Wide Range Achievement Test-Revised (Jastak & Wilkinson, 1984).
[j] California Achievement Test (CTB/Macmillan /McGraw Hill, 1992).
[k] Profile of Early Literacy Development (Dickinson & Chaney, 1998).
[l] Open Court (Hirshberg, Bereiter, & Hughes, 1989).
[m] Phonological Awareness Training for Reading (Torgesen & Bryant, 1994).
[n] Spelling Through Phonics (McCracken & McCracken, 1996).
[o] Wilson Reading System (Wilson, 1996).
[p] Test of Phonological Awareness-Early Elementary (Torgesen & Bryant, 1994).
[q] Woodcock Reading Mastery Test-Revised: Word Attack subtest (Woodcock, 1987).
[r] Letter-sound Correspondence Accuracy (Carnine, Silbert, & Kameenui, 1997).
[s] Invented Spelling (Tangel & Blachman, 1992).
[t] Code for Instructional Structure and Student Academic Response: Main Stream (Carta, Greenwood, Schulte, Arreaga-Mayer, & Terry, 1988).
[u] Informal Survey of Linguistic Knowledge (Moats, 1994; Moats & Lyon, 1996).
[v] Professional Development.

In contrast, we found that the more recent studies have consistently incorporated more of the five components into the professional development process, although perhaps as a consequence, comprehension received less emphasis. Seven of the nine studies targeted phonemic awareness skills (Baker & Smith, 1999; Bos, Mather, Narr, & Babur, 1999; Dickinson & McCabe, 2001; Jackson, Paratore, Chard, & Garnick, 1999; McCutchen et al, 2002; McCutchen & Berninger, 1999; O'Connor, 1999) and four addressed phonics and word identification skills (Bos et al., 1999; McCutchen et al., 2002; McCutchen & Berninger, 1999; O'Connor, 1999). Fluency was a focus in both studies conducted by McCutchen and Berninger (1999), McCutchen et al. (2002), and Jackson et al. (1999), and finally, vocabulary instruction was included in two investigations (Dickinson et al, 2001; O'Connor, 1999). Comprehension was only addressed by McCutchen & Berninger (1999) and McCutchen et al. (2002). It is noteworthy that only McCutchen et al. taught teachers to deliver all five components of reading instruction. Thus in terms of process, the more recent studies clearly and more consistently addressed the five components highlighted by the NRP.

Duration and intensity of professional development process. The NRP (2000) found a large variation in the length and intensity of professional development training. To illustrate the difficulty in comparing the duration of studies, consider the following two extremes. Coladarci and Gage (1984) never met with teachers and only provided them with a training packet for self-study. By contrast, in Conley's study training lasted 10–15 h. Furthermore, half of the studies reviewed by the NRP did not report the number of hours teachers were trained.

We continued to find in our review that a large variation in the duration of training for teachers, but at least researchers specified the period of time for initial formal training (ranging from two days to two and a half weeks). Encouragingly, several investigations also provided additional support, including meetings ranging from weekly to three throughout the intervention year, in-class mentoring with modeling and feedback, and on-going support and assistance from project staff. Further, researchers in the newer studies conceptualized professional development as occurring in better defined stages. Two teams enlisted teachers' help in defining their own needs (Baker & Smith, 1999; Dickson & Bursuck, 1999). We note that a few researchers used the summer months to provide intensive institutes that typically lasted two to two and a half weeks (Bos et al., 1999; McCutchen & Berninger, 1999; McCutchen et al., 2002; O'Connor, 1999). Only one research team created a University course for teachers (teachers received four credit hours) (Dickinson & McCabe, 2001).

These newer studies also provided more information about the frequency and intensity of a mentoring component, which again varied greatly across studies. In Baker and Smith (1999), the teacher was able to request mentoring from the research staff on an as needed basis. Bos et al. (1999) provided mentors who worked with individual teachers in their own classrooms on a monthly basis and O'Connor herself mentored teachers on a weekly basis. Dickson and Bursuck (1999) initiated bimonthly mentoring sessions during the second half of the year.

Unfortunately, only one study (O'Connor, 1999) "unpacked" or empirically tested any aspects of Showers et al.'s (1987) widely used principles of professional development. O' Connor manipulated the degree of support to directly compare the effects of low versus intense support. She found that teachers receiving the more intensive support had higher levels of implementation. However, as just mentioned, O'Connor herself provided this support, so without further replication with other "coaches", the generalizability of this finding is constrained. As we will discuss later, these differences make it virtually impossible to meaningfully compare the effectiveness of the professional development process across studies.

TEACHER PRESAGE AND TEACHER PRODUCT: TEACHER CHARACTERISTICS, KNOWLEDGE, AND BEHAVIOR

We have examined presage (a potential moderator variable) and teacher product (an outcome variable) jointly due to the fact that researchers have embraced a more dynamic view of teacher knowledge and are interested in its improvement. This shift reflects a need to know whether PD models can be linked to a measurable increase in teachers' knowledge about reading subject matter over time and/or an observed change in teachers' instruction. Other aspects of presage are more static and represent what individual characteristics teachers bring to their professional development experience that could conceivably vary and therefore moderate the effectiveness of professional development, including attitudes, knowledge, amount of education, experience, and even their cultural and ethnic background.

It is vital to learn whether the knowledge and skills presented during professional development is filtered through what the teachers bring to the experience. It is also critical to consider whether researchers have established equivalence between treatment and comparison conditions, particularly when teachers were not randomly assigned to condition and that researchers frequently contrasted volunteers to nonvolunteer controls (Vellutino & Schatschneider, 2004). Regrettably, only one of the seven quasi-experimental studies in the NRP (2000) review tested whether groups were equivalent prior to the studies, although several used pretest scores to control for possible differences in outcomes. Among the more recent studies we reviewed, the strongest evidence of control group equivalence was provided by McCutchen and colleagues (McCutchen & Berninger, 1999; McCutchen et al., 2002). These researchers matched schools on curriculum standards, instructional practices, student socioeconomic status (SES), and student ethnicity prior to assigning teachers to condition to increase the likelihood that treatment and comparison groups shared similar characteristics. The remaining eight recent studies utilized statistical methods to analyze whether groups differed significantly on target variables prior to intervention, which involved using pretest measures as covariates or in a repeated measures analysis of variance.

Generally speaking, across 13 NRP studies and 9 that we reviewed, investigators adequately described the teacher characteristics and demographics. For example, most described the educational levels of teachers (degrees earned), years of experience, and some provided their ethnicity. Of the studies reviewed by the NRP, only one quasi-experimental investigation (Miller, 1985) directly evaluated teachers' knowledge of the reading instruction prior to and after their participation in professional development using the *Inventory of Teacher Knowledge of Reading* (Artley & Hardin, 1975). Miller reported that controlling for pretest differences, participating teachers had higher levels of knowledge after professional development than did nonparticipating (i.e., control) teachers. Interestingly, on average, participating teachers who had volunteered for the study not only had higher levels of initial knowledge, but also had less experience and fewer advanced degrees than teacher who chose not to participate in professional development. Miller's findings suggest that the importance of presage variables should not be ruled out.

Our review located three additional recent investigations, which demonstrated that teachers receiving professional development increased their *knowledge* of phonology and linguistics significantly more than teachers who did not receive professional development (Bos et al., 1999; McCutchen & Berninger, 1999; McCutchen et al., 2002). One of these three also showed that teachers' *attitudes* toward explicit, structured language instruction improved following their participation in professional development (Bos et al., 1999). The reliability of these teacher measures is in question as a reliability coefficient was reported only for Bos et al.'s knowledge measure (it was adequate and Cronbach's alpha exceeded 0.80).

Most of the 13 studies in the NRP (2000) reviewed used *observations* to establish whether teachers implemented the strategies from the professional development. There was tremendous variation in the number of observations, from a minimum of two times occurring pre- and post-training (Coladarci & Gage, 1984; Streeter, 1986) to a maximum of more than 20 (Anderson, Evertson, & Brophy, 1979) occurring over a six-month period. Furthermore, most of the observation instruments were researcher-designed and only a small handful of the researchers provided reliability evidence.

Given the lack of psychometric information about the measures used, it may be helpful to at least consider the level of inference that is required of observers using these different instruments (for a more thorough description, see Haager, Gersten, Baker, & Graves, 2003). Low inference measurement systems (time sampling procedures, and classroom observation instruments specifically designed to measure targeted aspects of the literacy environment) require observers to follow clearly operationalyzed coding schemes for variables. By contrast, "high inference" observation instruments are typically more informal, subjective, and open-ended. Unfortunately, as we read the studies reviewed by NRP (2000), we found "high inference" was synonymous with simple global (yes/no) observations used to report whether teachers were implementing procedures.

Even in the more recent studies, classroom observational systems continued to vary in terms of their level of inference and in the frequency of observations. The number of teacher observations varied from monthly, to one or two times. Some of the recent studies used high inference methods, similar to those in earlier studies. However, we began to notice researchers using some more sophisticated low inference techniques such as computer-driven time sampling and formal observation instruments designed to capture specific behaviors, instructional techniques, or characteristics of the literacy environment. For example, the Code for Instructional Structure and Student Academic Response: Mainstream Version (MS-CISSAR) (Carta, Greenwood, Schulte, Arreaga-Mayer, & Terry, 1988) was used by Greenwood, Tapia, Abbott, and Walton (2003) as a low inference, computer-driven momentary time-sampling procedure. Every 20 s the observer was prompted to record an event from one of three categories (ecology, teacher, or student). The 30–60 min observations ($M = 39.8$) occurred during reading instruction with a mean percentage agreement of 97.4.

Notably, only one of the studies we reviewed, conducted by Dickinson and McCabe (2001), utilized a commercially available classroom observation tool with known psychometric properties. Test–retest reliability and interrater reliabilities exceeded 0.90 for the Early Language and Literacy Classroom Observation (ELLCO; Smith and Dickinson, 2002). Using this measure, the researchers found, teachers increased their use of instructional strategies focusing on evidence-based literacy practices. Teachers also made improvements in their classroom environments as a result of professional development. These changes resulted in significant between-group differences favoring teachers who received professional development in the provision of literacy areas and tools in the classroom including the presence and use of books, children's writing, and home support for literacy.

Furthermore, the fidelity of treatment was not consistently reported. Some studies provided general affirmations that teachers implemented "with fidelity", few provided the actual percentage of fidelity, and fewer still detailed inter-rater reliability of fidelity checks. Thus in terms of teacher presage and product, the more recent studies clearly and more consistently addressed the five components highlighted by the NRP. Evaluating the strength of linkage between professional development and teacher outcomes would be made easier if researchers would begin to (a) develop common measures that could be used across studies, (b) provide fidelity of implementation data, and (c) report the reliability of the measures.

Student Characteristics and Student Product

We found that the student SES and ethnic information was more consistently provided in the more recent studies we reviewed than in the older studies reviewed by the NRP (2000). We also noted that recent studies provided more information concerning whether or not the study population included students at-risk for or identified with reading disabilities or students who were English Language Learners. Four recently reviewed studies (Dickson & Bursuck, 1999; Greenwood et al., 2003; Jackson et al., 1999; O'Connor, 1999) provided disaggregated results of the student data for at-risk students.

The NRP (2000) reported that student change was measured using a wide variety of experimental measures, criterion-referenced, and norm-referenced measures and in our review, we also found the large range of measures, which made it impossible to directly compare student findings across studies. Nevertheless, like the NRP report, findings from our review suggest a trend in the data indicating that, on average, students significantly improved on most of the components they were taught. As discussed by the NRP, we also found few researchers provided effect size comparisons, which could assist in comparing treatment and control students' growth in a meaningful way across studies.

A potentially important improvement since the NRP-reviewed studies was in the type of student assessments used. The earlier studies typically used group administered assessments (i.e., the Gates–MacGinitie Reading Test, MacGinitie, MacGinitie, Maria, & Dreyer, 2000; the Stanford Achievement Test, The Psychological Corporation; California Achievement Test, CTB/Macmillan /McGraw Hill, 1992; Michigan Educational Assessment Program, Michigan Department of Education, 1994; or other high-stakes State achievement test). Sometimes these measures were administered only at the end of the school year, so researchers could not analyze potentially important differences in students' growth. In contrast, the more recent studies utilized more individually administered criterion-referenced tests (e.g., Dynamic Indicators of Basic Early Literacy Skills; DIBELS; Good & Kaminski, 2002) and norm-referenced tests (e.g., Test of Phonological Awareness, Torgesen & Bryant, 1994; Woodcock Reading Mastery Test – Revised, Woodcock, 1987; Peabody Picture Vocabulary Test, Dunn & Dunn, 1987). Criterion referenced measures allow a comparison to established benchmarks, and norm-referenced measures allow a comparison to standardized means, and national norms. Furthermore, as the newer generation of studies trained teachers in more of the five components of reading, they also more consistently measured student outcomes in these areas using multiple measures.

Unfortunately, it is still unclear whether professional development helped close the achievement gap. Of the four studies to provide disaggregated data or to test whether at-risk students caught up to typical peers, there were equivocal findings. O'Connor (1999) reported that students from professional development classrooms outperformed control students on measures of blending, segmentation, rapid letter naming, word identification, and dictation. Further, she found that children at risk (those with high incidence disabilities or pretest scores less then 85 on standardized measures of reading and spelling) in classrooms whose teachers received the more intense level of professional development did better on letter naming, word identification, and spelling. However, these children did not catch up to typical peers.

In the Dickson & Bursuck (1999) study, 11 out of 20 at-risk students made gains of 0.80 to 2.0 standard deviations on five out of six measures of reading after receiving small-group intensive instruction, but remained below the 25th percentile on oral reading fluency. In other instances, even though the students were reported to have benefited with growth in the desired direction, student gains were not significantly greater than children in the comparison group. For example, in Jackson et al. (1999) treatment students did make gains on measures of phonemic awareness and fluency, but did not catch up to their grade level peers. Greenwood et al. (2003) reported that high-risk children whose teachers got professional development made significant growth in oral reading fluency, but the slop and amount of their growth was lower than low risk children. Thus, on the basis of the existing research base, it is too early to conclude whether professional development is helping to close the achievement gap.

Context. We were unable to assess the impact of context as none of the studies in either review explored the potential moderating effect of school context on the effects of professional development. This is problematic given the data analytic methods used to date. However, four of the more recent studies (Dickinson & McCabe, 2001; Greenwood et al., 2003; McCutchen and Berninger, 1999; McCutchen et al, 2002) utilized more sophisticated methods of analysis that can more readily account for contextual variables: hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002). HLM allows researchers to account for the fact that student data exists within the organizational structure of the classroom and school. While these four studies used HLM, none of them used it to explore contextual variables.

DISCUSSION

This final section synthesizes what we have learned from the research and describes limitations. It also provides an exemplary study that links professional development to (a) improved teacher knowledge, attitude, and instruction and (b) improved student outcomes on multiple measures of the components taught during the professional development process. The section concludes with a discussion of implications for future research and practice.

What do we know based on current empirical research on professional development in beginning reading?

- We can identify a trend in the data across studies that suggests professional development efforts produced predominantly positive effects on teacher and student outcomes.
- On the other hand, if teachers were not seen to implement training or to change their teaching practices, students did not improve.

- Focus of more recent professional development consistently addresses scientifically based reading components.
- Recent studies have used more individual assessments and multiple measures to address student outcomes.
- Recent studies include more sophisticated data analytic procedures to address the nesting of students within teachers and schools and to examine growth over time.

Is it reasonable to expect general education classroom to be the first line of defense in efforts to prevent reading difficulties?

- It is too early to conclude, however, studies that included children at-risk for reading difficulties or children with disabilities reported substantial growth.
- We found no evidence that either children at-risk for reading difficulties or children with disabilities were able to catch up to their grade level peers.

What are some of the major limitations of the current database?

- None of the 22 studies reviewed could be considered a robust randomized experiment and therefore it is impossible to attribute causal links between teacher professional development and student outcomes.
- Few studies provided evidence of equivalent treatment and control groups.
- Only one study has evaluated the individual importance of particular aspects of professional development (O'Connor, 1999 examined the effect of mentoring).
- Several important moderator variables (presage, context) have not been explored.
- Only one study used a reliable and valid measure of teacher knowledge (Bos et al., 1999).
- The teacher observation measures were idiosyncratic, in other words they were so study-specific, that they lack generalizability.

Are there any examples of studies that provide linkage?

- Bos et al. (1999) established relatively clear and strong linkage between the professional development instructional topics, teacher measures, and student measures.
  - First, the focus of the professional development was to develop knowledge and skills in reading, spelling, and the structure of language.
  - Second, teachers' knowledge of the structure of language and their attitudes concerning early reading and spelling were reliably measured.

Findings indicated that the professional development teachers showed significant growth.

- Third, teachers were then observed in their classrooms to show they implemented the reading and spelling strategies presented during professional development with a high level of fidelity of implementation.
- Fourth, students were measured on multiple tests of reading, spelling, and phonological awareness. Students whose teachers received professional development had significantly greater gains on tests of sound identification, spelling of nonsense and real words, reading and spelling.

What suggestions could strengthen future research designs?

- The use of random assignment at either the teacher or the school level would allow for the strongest claims of causal inference. Randomization of teachers to conditions (within schools) is attractive, but could possible lead to "bleed-over" effects and possible conditions of compensatory rivalry from the teachers assigned to the control conditions. Randomization of schools to condition would minimize these threats to validity, but it is often difficult to find enough schools willing to participate.
- Future research designs could also focus on the components of professional development by creating treatment groups that systematically varied between who received or did not receive the various components.
- The development of reliable and valid measurement instruments aimed at assessing professional development would also increase our ability to link change in the teacher to changes in the classroom to changes in student outcomes.
- Systematic classroom observations would also aid in our understanding of the relationship between teacher classroom behaviors and student outcomes.
- Finally, child-level assessments that are conducted at pre- and posttest (at a minimum) should be sensitive enough across the school year to detect changes in the targeted areas of achievement.

Where do we go from here?

- Additional research is needed to examine the efficacy of professional development in schools serving children living in poverty and children who are from culturally diverse backgrounds.
- There is clearly a need for researchers to work together to develop more measures with adequate psychometric properties to address teacher knowledge and to adopt observational measures of instructional practice that have already been linked to improved student outcomes (e.g., *The*

English Language Learner Classroom Observation Instrument for Beginning Readers, Haager et al., 2003; *The Instructional Content Emphasis Instrument,* Edmonds & Briggs, 2003; *Classroom Language Arts Systematic Sampling and Instructional Coding Observation Systems,* Scanlon, Gelzheiser, Fanuele, Sweeney, & Newcomer, 2003).

- Specifically at this stage of inquiry, it could be helpful to include multiple measures – some high and some low inference. Low inference measures are needed to document the fidelity with which teachers implement what they are taught to do, but more elaborate field notes or technology supporting coding systems could offer important information about how well teachers tailor instruction for individual children, and might capture other dimensions of quality we do not yet know about.
- As Zeichner (2005) recently pointed out in his discussion of a research agenda related to teacher education we need a ''chain of inquiry around particular questions and consistently defined outcomes, and of researchers using the same outcome measures across studies''(p. 742). This agenda should include collaboration across sites and replication of findings, as well as a systematic ''unpacking'' of the professional development process.

REFERENCES

Adams, M. J. (1990). *Learning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Allington, R. L. (2002). *Big brother and the national reading curriculum: How ideology trumped evidence*. Portsmith, NH: Heineman.

Anderson, L. M., Evertson, C. M., & Brophy, J. E. (1979). An experimental study of effective teaching in first-grade reading groups. *Elementary School Journal*, *79*(4), 193–223.

Artley, A. S., & Hardin, V. (1975). *Inventory of teacher knowledge of reading: Information and answer key*. Columbia: University of Missouri.

Baker, J. E. (1977). Applications of the in-service training/classroom consultation model to reading instruction. *Ontario Psychologist*, *9*(4), 57–62.

Baker, S., & Smith, S. (1999). Starting off on the right foot: The influence of four principles of professional development in improving literacy instruction in two kindergarten programs. *Learning Disabilities Research & Practice*, *14*(4), 239–253.

Bos, C. S., Mather, N., Narr, R. F., & Babur, N. (1999). Interactive, collaborative professional development in early literacy instruction: Supporting the balancing act. *Learning Disabilities Research & Practice*, *14*(4), 227–238.

Brady, S., & Moats, L. (1997). Informed instruction for reading success: Foundations to teacher preparation. *A position paper of the International Dyslexia Association.*

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal*, *10*, 245–252.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In: M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 340–370). NY: Macmillan.

Brown, R., Pressley, M., Van Meter, P., & Schuder, T. (1996). A quasi-experimental validation of transactional strategies instruction with low-achieving second-grade readers. *Journal of Educational Psychology*, *88*(1), 18–37.

California Achievement Test. (1992). CTB/Macmillan /McGraw Hill.

Carnine, D. W., Silbert, J., & Kameenui, E. J. (1997). *Direct instruction reading* (3rd ed.). Columbus, OH: Merrill.

Carta, J. J., Greenwood, C. R., Schulte, D., Arreaga-Mayer, C., & Terry, B. (1988). *Code for instructional structure and student academic response: Mainstream version (MS-CISSAR)*. Kansas City: University of Kansas, Bureau of Child Research, Juniper Gardens Children's Project.

Clancy-Menchetti, J., & Al Otaiba, S. (2006). *The effectiveness of professional development*. Manuscript in preparation.

Clay, M. (1985). *The early detection of reading difficulties* (3rd ed.). Portsmouth, NH: Heinemann.

Cochran-Smith, M., & Lyttle, S. (1999). Relationships of knowledge and practice: Teacher learning in communities. In: A. Iran-Nejar & P. D. Pearson (Eds), *Review of Research in Education (24)* (pp. 249–305). Washington DC: AERA.

Cochran- Smith, M., & Zeichner, K. M. (2005). *Studying teacher education: The report of the AERA Panel on research and teacher education*. Mahwah, NJ: Erlbaum.

Coladarci, T., & Gage, N. L. (1984). Effects of a minimal intervention on teacher behavior and student achievement. *American Educational Research Journal*, *21*(3), 539–555.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. US Government Printing Office: Washington, DC.

Conley, M. M. W. (1983). Increasing students' reading achievement via teacher inservice education. *Reading Teacher*, *36*(8), 804–808.

Darling-Hammond, L. (2000a). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives*, *8*(1) http://epaa.asu.edu/epaa/v8nl.

Deford, D. E. (1985). Validating the construct of theoretical orientation in reading instruction. *Reading Research Quarterly*, *20*, 351–367.

Dickinson, D. K., & Chaney, C. (1998). *Profile in Early Literacy Development*. Newton, MA: Education Development Center, Inc.

Dickinson, D. K., & McCabe, A. (2001). Bringing it all together: The multiple origins, skills, and environmental supports of early literacy. *Learning Disabilities Research & Practice*, *16*(4), 186–202.

Dickson, S. V., & Bursuck, W. D. (1999). Implementing a model for preventing reading failure: A report from the field. *Learning Disabilities Research & Practice*, *14*(4), 191–202.

Duffy, G. G., Roehler, L., Meloth, M. S., Vavrus, L. G., Wesselman, R., Putnam, J., & Bassiri, D. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly*, *21*(3), 237–252.

Duffy, G. G., Roehler, L., Sivan, E., Rackliffe, G., Book, C., Meloth, M. S., Vavrus, L. G., Wesselman, R., Putnam, J., & Bassiri, D. (1987). Effects of explaining the reasoning associated with using reading strategies. *Reading Research Quarterly*, *22*(3), 3347–3368.

Dunkin, M., & Biddle, B. (1974). *The study of teaching*. New York: Holt, Rinehart & Winston.

Dunn, L. M., & Dunn, L. M. (1987). *Peabody Picture Vocabulary TestThird Edition*. Circle Pines, MN: American Guidance Service.

Edmonds, M., & Briggs, K. L. (2003). The instructional content emphasis instrument: Observations of reading instruction. In: S. Vaughn & K. L. Briggs (Eds), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 31–52). Baltimore, MD: Brookes.

Foorman, B. R., Francis, D. J., Winikates, D., Mehta, P., Schatschneider, C., & Fletcher, J. M. (1997). Early interventions for children with disabilities. *Scientific Studies of Reading*, *1*, 255–276.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice*, *13*(4), 204–219.

Gardner, E. F. (1982). *Stanford Achievement Test*. New York: The Psychological Corporation, Harcourt Brace.

Gilmore, J., & Gilmore, E. (1968). *The Gilmore oral reading test*. San Antonio, TX: Harcourt Brace Educational Measurement.

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Retrieved September 1, 2005, from http://www.urban.org/UploadedPDF/410958_NBPTSOutcomes.pdf

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*, 257–288.

Good, R. H., & Kaminski, R. A. (Eds). (2002). Dynamic indicators of basic early literacy skills (6th ed.). Eugene, OR: Institute for Development of Educational Achievement.

Greenwood, C. R., Tapia, Y., Abbott, M., & Walton, C. (2003). A building-based case study of evidence-based literacy practices: Implementation, reading behavior, and growth in reading fluency, k-4. *The Journal of Special Education*, *37*(2), 95–110.

Grossman, P. L. (1990). A study in contrast: Sources of pedagogical content knowledge for secondary English. *Journal of Teacher Education*, *40*(5), 24–31.

Grossman, P. L. (1991). What are we talking about anyway? Subject-matter knowledge of secondary English teachers. In: J. Brophy (Ed.), *Advances in research on teaching* (Vol. 2, pp. 245–264). Greenwich, CT: JAI Press.

Haager, D., Gersten, R., Baker, S., & Graves, A. W. (2003). The English-language learner classroom observation instrument for beginning readers. In: S. Vaughn & K. L. Briggs (Eds), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 111–144). Brookes Publishers.

Heinesman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Review*, *51*, 358–374.

Hirshberg, J., Bereiter, C., & Hughes, A. (1989). *Open court*. LaSalle, IL: Open Court.

Hoffman, J. V. (1991). Teacher and school effects in learning to read. In: R. Barr, M. L. Kamil, P. B. Mosenthal & P. D. Pearson (Eds), *Handbook of reading research*, (Vol. II, pp. 911–950). New York: Longman.

Jackson, J. B., Paratore, J. R., Chard, D. J., & Garnick, S. (1999). An early intervention supporting the literacy learning of children experiencing substantial difficulty. *Learning Disabilities Research & Practice*, *14*(4), 254–267.

Jastak, S., & Wilkinson, G. S. (1984). *Wide Range Achievement – Revised (WRAT–R)*. Wilmington, DE: Jastak Associates.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*, 437–447.

Kaminski, R. A., & Good, R. H. (1996). Toward a technology of assessing basic early literacy skills. *School Psychology Review*, *25*, 215–227.

Kennedy, M. M., Birman, M., & Demaline, B. (1986). *The effectiveness of Chapter 1 services. Second interim report from the National Assessment of Chapter 1*. Washington, DC: Office of Educational Research and Improvement (OERI).

Knapp, M. S. (Ed.) (1995). *Teaching for meaning in high-poverty classrooms*. New York: Teachers College Press.

Lewis, L., Parsad, B., Carey, N., Bartfai, N., Farris, E., & Smerdon, B. (1999). *Teacher quality: A report on the preparation and qualifications of public school teachers*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Lyon, G. R. (1985). Identification and remediation of learning disability subtypes: Preliminary findings. *Learning Disability Focus*, *1*, 21–35.

MacGinitie, W. H., MacGinitie, R. K., Maria, R., & Dreyer, L. G. (2000). *Gates–MacGinitie Reading Tests (GMRT)*. Itasca, IL: Riverside Publishing a Houghton Mifflin Company.

Mastropieri, M. A., Leinart, A., & Scruggs, T. E. (1999). Strategies to increase reading fluency. *Intervention in School and Clinic*, *34*(5), 278–292.

Mather, N., Bos, C., & Babur, N. (2001). Perceptions and knowledge of preservice and inservice teachers about early literacy instruction. *Journal of Learning Disabilities*, *34*, 472–482.

McCracken, M., & McCracken, R. (1996). *Spelling through phonics*. Peguis Publishers.

McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, N., Cox, S., Potter, N. S., Quiroga, T., & Gray, A. L. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities*, *35*, 69–86.

McCutchen, D., & Berninger, V. W. (1999). Those who know, teach well: Helping teachers master literacy-related subject-matter knowledge. *Learning Disabilities Research & Practice*, *14*, 215–226.

Michigan Educational Assessment Program (MEAP). (1994). Michigan Department of Education.

Miller, J. W., & Ellsworth, R. (1985). The evaluation of a two-year program to improve teacher effectiveness in reading instruction. *Elementary School Journal*, *85*, 485–496.

Moats, L. C. (1994). The missing foundation in teacher education: Knowledge of the structure of spoken and written language. *Annals of Dyslexia*, *44*, 81–102.

Moats, L. C. (2000). *Speech to print: Language essentials for teachers*. Baltimore, MD: Brookes.

Moats, L. C., & Lyon, G. R. (1996). Wanted: Teachers with knowledge of language. *Topics in Language Disorders*, *16*, 73–81.

National Assessment of Educational Progress. (2003). Retrieved July 8, 2005, from http://nces.ed.gov/nationsreportcard/

National Center for Educational Statistics. (2002). *The Nation's Report Card, 2002 National Assessment of Educational Progress.* Retrieved November 26, 2002, from http://nces.ed.gov/nationsreportcard/

National Reading Panel (NRP). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, (H. R. 1).

O'Connor, R. E. (1999). Teachers learning ladders to literacy. *Learning Disabilities Research & Practice*, *14*, 203–214.

Office of Educational Research and Improvement [OERI]. (1997). *National Awards Program for model professional development 1998 application*. Washington, DC: Author.

Phelps, G., & Schilling, S. (in press). Developing measures of content knowledge for teaching reading. *Elementary School Journal*.

President's Commission on Excellence in Special Education. (2002). *A new era: Revitalizing special education for children and their families*. Washington, DC: Author.

Pressley, M., Wharton-Mc Donald, R., Allington, R., Block, C. C., Morrow, L., Tracey, D., Baker, K., Brooks, G., Cronin, J., Nelson, E., & Woo, D. (2001). A study of effective first-grade instruction. *Scientific Study of Reading*, *5*, 35–58.

Puma, M. J., Karweit, N., Price, C., Ricciuti, A., Thompson, W., & Vaden-Kiernan, M. (1997). *Prospects: Final report on student outcomes*. Washington, DC: Planning and Evaluating Service, U.S. Department of Education.

Rivers, J. C., & Sanders, W. L. (2002). Teacher quality and equity in educational opportunity: Findings and policy implications. In: L. T. Izumi & W. M. Eders (Eds), *Teacher quality* (pp. 13–24). Stanford, CA: Hoover Institution Press.

Rosenshine, B. (1979). Content, time, and direct instruction. In: P. Peterson & H. Walberg (Eds), *Research on teaching: Concepts, findings, and implications* (pp. 28–56). Berkeley, CA: McCutchan.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In: R. M. W. Travers (Ed.), *Second handbook of research on teaching* (pp. 122–183). Chicago: Rand McNally.

Scanlon, D., Gelzheiser, L., Fanuele, D., Sweeney, J., & Newcomer, L. (2003). *Classroom language arts systematic sampling and instructional coding observation systems (CLASSIC)*. Albany: Child Research and Study Center.

Schereens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon Press.

Schonell, F. J., & Schonell, P. E. (1960). *Schonell graded word reading test (SGWRT)*. Diagnostic and Attainment Testing. Edinburgh: Oliver and Boyd.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized casual inference*. Boston: Houghton-Mifflin.

Shavelson, R. J., & Towne, L. (2002). *Scientific research in education*. Washington DC: National Academy Press.

Showers, B., Joyce, B., & Bennett, B. (1987). Synthesis of research on staff development: A framework for future study and a state-of-the-art analysis. *Educational Leadership*, *45*(3), 77–87.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Smith, M. W., & Dickinson, D. K. (2002). *Early language and literacy classroom observation (ELLCO)*. Education Development Center, Inc., Newton, MA: Brookes Publishing.

Snow, C. E. (2002). *Reading for understanding: Toward an R & D program for comprehension*. Santa Monica, CA: RAND.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds) (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Specific Learning Disabilities: Finding Common Ground. (2002). Washington, DC: U.S. Department of Education, Office of Special Education Programs.

Spira, E. G., Bracken, S. S., & Fischel, J. E. (2005). Predicting improvement after first grade reading difficulties: The effects of oral language, emergent literacy, and behavior skills. *Developmental Psychology*, *41*, 225–234.

Stallings, J., & Krasavage, E. M. (1986). Program implementation and student achievement in a four-year Madeline Hunter follow-through project. *Elementary School Journal*, *87*, 117–138.

Stallings, J., Robbins, P., Presbrey, L., & Scott, J. (1986). Effects of instruction based on the Madeline Hunter model on students' achievement: Findings from a follow-through project. *Elementary School Journal*, *86*, 571–587.

Stanford Achievement Test. San Antonio, TX: The Psychological Corporation.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360–407.

Streeter, B. B. (1986). The effects of training experienced teachers in enthusiasm on students' attitudes toward reading. *Reading Psychology*, *74*, 249–359.

Sweet, A. P., & Snow, C. E. (Eds) (2003). *Rethinking reading comprehension (Solving problems in teaching of literacy)*. New York, NY: Guildford Press.

Talmage, H., Pascarella, E. T., & Ford, S. (1984). The influence of cooperation learning strategies on teacher practices, student perceptions of the learning environment, and academic achievement. *American Educational Research Journal*, *21*, 163–179.

Tangel, D. M., & Blachman, B. (1992). Effect of phoneme awareness instruction on kindergarten children's invented spelling. *Journal of Reading Behavior*, *24*, 233–261.

Taylor, B. M., Pearson, P. D., Clark, K., & Walpole, S. (1999). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. In: B. M. Taylor & P. D. Pearson (Eds), *Teaching reading: Effective schools, accomplished teachers* (pp. 3–72). Mahwah, NJ: Erlbaum.

Taylor, B. M., Pressley, M., & Pearson, P. D. (2002). Research supported characteristics of teachers and schools that promote reading achievement. In: B. M. Taylor & P. D. Pearson (Eds), *Teaching reading: Effective schools, accomplished teachers* (pp. 361–374). Mahwah, NJ: Erlabaum.

Torgesen, J. K., & Bryant, B. R. (1994). *Phonological awareness training for reading*. Austin, TX: PRO-ED.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying with reading/learning disabilities. *Exceptional Children*, *69*, 391–409.

Vellutino, F. R., & Schatschneider, C. (2004). Experimental and quasi-experimental design in literacy research. In: N. K. Duke & M. H. Mallette (Eds), *Literacy Research Methodologies*. New York: Guilford.

Venezky, R. L., & Winfield, L. (1979). *Schools that succeed beyond expectation in teaching reading* Technical Report No.1. Newark, DE: Department of Education Studies, University of Delaware.

Weber, G. (1971). *Inner city children can be taught to read: Four successful schools (CGE Occasional Papers No. 18; ERIC Document Reproduction Service No. Ed 057 125).* Washington, DC: Council for Basic Education.

Wharton-McDonald, R., Pressley, M., & Hampston, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and student achievement. *The Elementary School Journal*, *99*(2), 101–128.

Whitehurst, G. (2002). *Scientifically based research on teacher quality: Research on teacher preparation and professional development*. Paper presented at the White House Conference on Preparing Tomorrow's Teachers, Washington, DC.

Wilder, G. (1977). Five exemplary reading programs. In: J. T. Guthrie (Ed.), *Cognition, curriculum, and comprehension* (pp. 57–68). Newark, DE: International Reading Association.

Wilson, B. A. (1996). *Wilson Reading System*. Austin, TX: PRO-ED, Inc.

Woodcock, R. (1987). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *WJIII Tests of cognitive abilities and achievement*. Itasca, IL: Riverside Publishing.

Yopp, H. K. (1995). A test for assessing phonemic awareness in young children. *The Reading Teacher*, *49*, 20–29.

Zeichner, K. M. (2005). Research agenda for teacher education. In: M. Cochran-Smith & K. M. Zeichner (Eds), *Studying teacher education: The report of the AERA Panel on research and teacher education* (pp. 737–760). Mahwah, NJ: Erlbaum.

READING COMPREHENSION AND WRITTEN COMPOSITION PROBLEMS OF CHILDREN WITH ADHD: DISCUSSION OF RESEARCH AND METHODOLOGICAL CONSIDERATIONS

Ana Miranda, Manuel Soriano and Rosa García

ABSTRACT

The present study analyzed the performance of children with attention deficit hyperactivity disorder (ADHD) when carrying out reading comprehension and written composition tasks. Thirty children with ADHD and 30 normally developing children without ADHD, matched on age, IQ, word retrieval and spelling, were selected. All of the subjects were evaluated using four types of reading comprehension tasks (literal comprehension, inferential comprehension, a fragment ordering task, and recall of story content), and a composition writing task. The results indicate that the two groups (ADHD and without ADHD) do not differ on literal comprehension or inferential comprehension. Nevertheless, our results show that children with ADHD perform significantly worse than the group without ADHD on the fragment ordering task, the recall of story content,

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 237–256

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19009-6

and on different indicators of written language production, which depend primarily on self-regulation abilities necessary for organizing information and maintaining the level of effort. The findings suggest that the deficit observed in reading comprehension and written composition skills in children with ADHD may reflect deficiencies in executive processes. The methodology used in this research on the reading comprehension and written composition problems of children with ADHD presents a series of strengths and weaknesses. The reflections on the limitations identified in the study serve as a basis for establishing directions for future research.

Students with attention deficit hyperactivity disorder (ADHD) usually suffer from academic underachievement, which may be largely due to their problems with self-regulation, a system that essentially includes three components: an attentional component, an inhibitory component, and a strategic and organizational component (Douglas, 2005). Furthermore, the long-term academic functioning of children with ADHD is worse when the behavioral problems are more severe (Barry, Lyman, & Klinger, 2002) and when the disorder is associated early on with learning difficulties (Faraone, Bierderman, Monuteaux, Doyle, & Seidman, 2001).

Approximately 70% of the children with ADHD present some kind of learning difficulty (Mayes, Calhoun, & Crowell, 2000). Specifically, ADHD and reading disorder (RD) co-occur significantly more frequently than would be expected by chance, as the proportion of RD in samples of children with ADHD falls between 25 and 40% (Semrud-Clikeman et al., 1992). The oscillation of the percentages provided on the comorbidity of the two disorders, RD and ADHD, is basically due to the diagnostic criteria used in the selection of the samples, the control of intelligence levels and the type of reading examined.

RD is a heterogeneous disorder that can affect not only the basic decoding skills but also the ability to derive meaning from a text. However, in contrast with the numerous studies that have analyzed possible problems in word reading skills in students with ADHD, very few studies have dealt with the analysis of the reading comprehension difficulties they might experience. However, there are compelling reasons why this matter needs to be studied, as reading comprehension not only directly affects academic success but it is also a skill that is important for the individual's progress throughout his lifetime. In this sense, the results of recent studies suggest that, at least as a group, adolescents and adults with ADHD present lower performance levels on text comprehension tasks. Thus, Ghelani, Sidhu, Jain, and

Tannock (2004) have reported that adolescents with ADHD exhibit adequate single word reading abilities, but they experience subtle difficulties on measures of text reading rate and silent reading comprehension. Samuelson, Lundberg, and Herkner (2004) also reported that adults with ADHD had lower results on text comprehension tasks, although they did not find significant differences between adults with and without ADHD on word recognition or phonological skill tasks.

Previous studies that have analyzed the comprehension of stories by children with ADHD have been based on the oral narration of these stories. In general terms, the studies carried out found that children with ADHD did not show difficulties in understanding or identifying the main idea of the narrations (Tannock, Purvis, & Schachar, 1993; O′Neill & Douglas, 1991; Zentall, 1988). However, when more elaborate indicators were used in the analysis of the production of the stories (basically stories that were visualized or heard earlier), the results were much richer. Thus, different studies (Lorch et al., 2004b; Renz et al., 2003; Lorch et al., 1999; Purvis & Tannock, 1997) have encountered two consistent findings: (a) the narrations by the children with ADHD differ from those of normal children in aspects related to the structure of the story, as they refer less to achieving the goal, which is the main point of the complete representation of the story; (b) the children with ADHD made fewer causal connections between ideas/unit, and they showed less thematic sensitivity. Furthermore, the coherence of the narration of the children with ADHD was affected by a greater number of errors (e.g. ambiguous references, embellishing, ... etc.) than was the case of the children in the control group.

On the other hand, several studies have provided conclusions useful for better understanding factors related to the task itself that might be influencing the performance of children with ADHD on comprehension tasks. In this framework, the results of Brock and Knapp (1996) highlight the difficulties children with ADHD have in identifying the main ideas in expository texts, in spite of having average word reading skills. Furthermore, the performance of children with ADHD without added linguistic difficulties seems to worsen as the text gets longer (Cherkes-Julkowski, Stolzenberg, Hatzes, & Madaus, 1995). Likewise, the reading mode also seems to influence performance on reading comprehension tests, as in ADHD adolescents there is a tendency toward lower scores on silent reading comprehension tasks (Gelani, Sidhu, Jain, & Tannock, 2004).

In contrast with the advances made in the study of reading comprehension with students with ADHD, we do not know very much about their performance on writing tasks. In a study carried out by Resta and Eliot (1994), the control group scored significantly higher than the students with ADHD on the overall test composite, on general writing ability, on word complexity and on productivity scores (i.e., they wrote more words). On the other hand, Ross, Poidevant, and Miner (1995) did not find group differences regarding the quantity of letters produced per minute of students with attention and hyperactivity problems or confirmed ADHD diagnoses and the control group. Therefore, the results from the two studies cited above point out that: (a) students with attention deficits can write at the same speed as children without ADHD; and (b) students with ADHD are less proficient writers (having lower overall composite scores), and they do not produce texts that contain as many words as their normally achieving peers do. However, it is obvious that the studies carried out so far have not provided an adequate analysis of the writing disabilities of students with ADHD.

With these considerations in mind, the purpose of the present study is to advance the study of the characteristics of children with ADHD when performing comprehension and text writing tasks. Specifically, we proposed two objectives: (a) to analyze the possible reading comprehension problems of children with ADHD who do not have word recognition deficits, using four tasks with different processing demands: literal comprehension, inferential comprehension, ordering fragments and recall after reading a story; and (b) to investigate the possible text writing problems of children with ADHD who do not have problems writing words.

METHOD

Subjects

The participants were 60 students selected for the experimental and control groups. Most of the children were from families with a low socioeconomic status, but without cultural or environmental disadvantages. All of the subjects were Caucasian and spoke Spanish as their primary language. They ranged in age from 7 to 12 years 1 month ($M = 9$ years, 1 month; $SD = 1$ year, 9 months). They were students in primary education from different schools in Valencia (Spain).

Of the entire sample, 30 children had ADHD, while the remaining 30 made up the control group of children without ADHD. The 30 children with ADHD had been diagnosed in the Neuropediatric Service of the Children's Hospital "La Fe" in Valencia, according to the following criteria: (a) a total score of 12 or more on the questionnaire for parents and teachers from the

DSM-IV (American Psychiatric Association, 1994), on the sections on Inattention and Hyperactivity-Impulsiveness, (b) the duration of the symptoms was greater than one year, (c) the problem had appeared before the age of 7, (d) an IQ score of 80 or more as measured by the WISC-R (Wechsler, 1993), (e) absence of psychosis, neurological damage or sensory or motor deficits and (f) none of the children were receiving pharmacological treatment.

To obtain the control group without ADHD (30 subjects), the collaboration of the School Psycho-pedagogical Services was requested. The criteria that were adopted to select the children were: (a) normal academic progress; (b) an IQ score above 80 on the WISC-R; and (c) absence of language or learning problems, psychosis or neurological damage, sensory deficits or motor deficits.

Specifically, of the 60 children who made up the entire sample, 47 were boys and 13 girls, representing 78.3 and 21.6%, respectively. Likewise, of the 30 children who made up the group with ADHD, 90 were boys and 10% girls, and of the 30 children in the control-normal group, 66.7 were boys and 33.3% girls.

In addition, as reflected by the results in Table 1, the children in both groups (with and without ADHD) were balanced on Verbal IQ and Vocabulary (WISC-R; Wechsler, 1993), as well as on word recognition and word writing errors (TALE; Cervera & Toro, 1984).

Table 1. Age, Verbal IQ, Vocabulary, Reading and Writing Scores for ADHD and without ADHD Control Children.

	Without ADHD ($N = 30$)		With ADHD ($N = 30$)			
	Mean (DT)	Average (IQR)	Mean (DT)	Average (IQR)	Statistical *F*/*U*	Signif. Bilat.
Age	9.10 (1.90)	9 (4)	9.10 (1.90)	9 (4)	$U = 450$	1
Verbal IQ	102.67 (11.29)	–	94.23 (12.10)	–	$F = 2.03$	0.162
Vocabulary	30.47 (17.90)	26.50 (24.25)	26.43 (9.69)	26.50 (18.25)	$U = 408.5$	0.54
Word reading errors	2.43 (4.02)	2.00 (1.50)	3.77 (4.14)	2.00 (5.50)	$U = 377$	0.272
Word writing errors (or spelling errors)	4.10 (4.50)	3.00 (6.50)	3.43 (2.39)	3.00 (4.00)	$U = 421.5$	0.670

Note: The data are presented as mean/standard deviation, in the case of a normal distribution and/or as interquartilic average/range (IQR) in the opposite case.

Evaluation Instruments and Procedure

Following the objectives of this study, different comprehension and text composition tasks were chosen. All of the tasks were administered to the children individually during three sessions by one of the authors in a noise-free room.

1. *Comprehension processes*. Four measures of text comprehension were used that presented different cognitive demands: literal comprehension, inferential comprehension, fragment ordering and story recall.

(a) *Literal comprehension* was evaluated by means of a comprehension subtest (TALE; Cervera & Toro, 1984). The children had to read a narrative or descriptive text only once and then answer 10 open-ended literal questions about the text.

(b) *Inferential comprehension.* In this task the children had to read a story and answer questions of an inferential nature (Gárate, 1994). The stories followed the categories from the story grammar by Stein and Glenn (1979). The first story, "Luis and his Bicycle", which was used by 7–9-year-old children, was composed of one episode and 20 propositions. The second story, "The Shepherd and the Dwarf", was made up of two episodes and 28 propositions, and it was used with children from 10 to 12 years of age. All of the questions presented (7 in the first story and 8 in the second) required making anaphoric inferences or bridge inferences. In the correction, one point was given for the answers that were correct from a semantic point of view, half a point for the incomplete or partially correct answers, and zero points for the incorrect answers.

(c) *Ordering fragments.* In this task the children had to order fragments of the story "Juan and his Balloon" (Paniagua, 1983). The 7–9 year olds were presented with the first episode (6 fragments), and both episodes were presented to the 10–12 year olds (11 fragments). All the subjects were given and read the first fragment, and they were asked to put the rest in order: "*Now you have to put these cards in order so that they tell a nice story*". From the tasks, two scores were obtained: the accuracy of the ordering and the number of movements used in this ordering. The accuracy of the ordering was evaluated by giving one point for each fragment organized correctly, that is, preceded and followed by the corresponding fragment; half a point was given when the fragment was preceded or followed by another with which there was some kind of syntactic and/or semantic connection; and the score was 0 in any other situation. On the other hand, in the second measure related to the number of movements, the number of times the children moved the cards in order to organize them was counted.

(d) *Story recall.* In this task the children had to tell a story they had read before. Two stories were used that followed the categories from the story grammar by Stein and Glenn (1979) and constructed based on stories used by Gárate (1994). The first of the stories, "María and her Duck", made up of 20 propositions and a simple structure (simple syntax, one episode, only one character and the action taking place in one limited setting-time framework), was used with children with a chronological age from 7 to 9. The second story, "The Boy and the Genie", made up of 28 propositions and two episodes and presenting a much more complex causal network (literary embellishments, more than one character), was used with the children with a chronological age from 10 to 12 years. The stories were presented in written form, due to the fact that students with ADHD are more effective on tasks with a visual modality (Webster, Hall, Brown, & Bolen, 1996), and the oral reading of the stories was selected because silent reading reduces comprehension in children with ADHD (Dubey & O'Leary, 1975).

The evaluator gave the children the following instructions "*Now you are going to read a story aloud. Read it carefully because later you will have to tell it to me.*" The children's narration was taped on a cassette, with the evaluator adopting a neutral attitude during the taping, so that his or her interventions would not help the children. The tapes of the subjects' retelling of the story were transcribed and evaluated by four people with Psychology degrees who had been previously trained. An interrater agreement that varied between 87 and 91% was obtained for the different variables used in the analyses.

The global production measures for the narration analyzed were: (a) *Proportion of the total number of propositions recalled.* This was calculated by counting the total number of propositions that respected the semantic content of the proposition, not necessarily its literal meaning (e.g. "He left the duck on top of a rock"), dividing it by the total number of propositions in each story and multiplying this by 100 and (b) *Proportion of propositions by categories, or number of propositions recalled in each of the categories by* Stein and Glenn (1979), that is, introduction, event, internal response, action, outcome and resolution, divided by the total number of propositions in each of them and multiplying the results by one hundred (See Appendix 1).

2. *Text composition processes.* Here a spontaneous writing task was used (TALE; Cervera & Toro, 1984), which consisted of asking the children to write an essay about an excursion. They were given the following

instructions: "*Now we are going to write an essay. You have to write here everything you can think of about a trip you took.*" The children's essays were evaluated using different traditional measures for evaluating written discourse (Cervera & Toro, 1984; Calsamiglia & Tusón, 1999): (a) *Time* used by each child in completing the task; (b) *the number of words written*, which was derived by counting the total number of words written in each writing sample; (c) *the number of sentences*, characterized by subject, verb, predicate, and grammatically identifiable relationships (Alarcos Lorach, 1995); no distinction was made between simple sentences with coordinate clauses and compound sentences with subordinate clauses; (d) *Mean Length of Sentences* (Clemente, 1989), which was obtained following the method of counting words, by dividing the total number of words by the number of sentences (MLSp); (e) *Syntactic Errors* in the use of number, gender, verb tenses, word order in the sentences, omissions, substitutions or addition of functional words (e.g., articles, prepositions, adverbs), the incoherence of the text, as well as the enumeration of words without syntactic agreement or the repetition of sentences; (f) *Expressive Content,* where the number of verbs, adjectives, adverbs were counted, as well as the conjunctions and phrases in the sentences that express cause-effect relationships; and (g) *Type-Token Ratio,* which was obtained by dividing the total number of different words by the total number of words, following the criteria adapted to Spanish by Clemente (1995) and Serra, Serrat, Bel, and Aparici (2000).

The compositions were evaluated by four people who held degrees in psychology and had been previously trained. An interrater agreement between 85 and 100% was obtained on the different variables used in the analysis.

RESULTS

Before performing the statistical analyses designed to compare the possible differences between the control group and the ADHD group, the data was shown to meet the criterion of statistical normalcy by applying the Shapiro–Wilks test. In those cases where the distribution was normal ($p > 0.05$), an analysis of variance (ANOVA) of comparison between groups was performed, while the Mann–Whitney U test was used in the opposite case. A p bilateral value inferior to 0.05 was defined as significant.

Results of the text comprehension measures. The results presented in Table 2 indicate that the group of children with ADHD and the controls did not differ on literal comprehension, U (58) = 351.5, $p < 0.141$, or inferential comprehension, U (58) = 440, $p < 0.881$. On the other hand, the children with ADHD showed a significantly worse performance than the control group on the accuracy with which they carried out the fragment ordering task, $U\,(58) = 302, p < 0.022$. However, in the number of movements used to order the fragments, no significant differences were found between the group with ADHD and the group without ADHD, $U\,(58) = 336, p < 0.088$.

The results related to the global measures of narrative production,[1] which appear in Table 2, indicate that the children with ADHD and the normal children differ with regard to the number of propositions recalled, $F(1{,}58) = 5.991$, $p < 0.017$. Furthermore, regarding the analysis carried out

Table 2. Comparison of Children with ADHD and Children without ADHD on the Four Measures of Text Comprehension.

	Without ADHD ($N = 30$)		With ADHD ($N = 30$)			
	Mean (DT)	Average (IQR)	Mean (DT)	Average (IQR)	Statistical *F*/*U*	Signif. Bilat.
Literal comprehension	7.83 (2.10)	8.00 (2.25)	6.82 (2.49)	7.00 (4.00)	$U = 351.5$	0.141
Inferential comprehension	4.70 (1.49)	4.75 (3.00)	4.63 (1.47)	5.00 (2.12)	$U = 440$	0.881
Fragment ordering accuracy	5.50 (2.33)	4.00 (5.00)	4.05 (2.69)	4.00 (4.12)	$U = 302$	0.022
Fragment ordering number of movements	7.97 (3.24)	7.00 (5.00)	9.33 (3.58)	8.50 (6.25)	$U = 336$	0.088
% of propositions	62.07 (15.19)	–	50.14 (21.94)	–	$F = 5.991$	0.017
Introduction	63.33 (38.13)	75.00 (56.25)	41.66 (39.57)	25.00 (81.25)	$U = 319$	0.046
Event	78.33 (20.48)	75.00 (50.00)	60.00 (30.52)	62.50 (25.00)	$U = 297$	0.018
Internal response	36.67 (34.57)	50.00 (50.00)	33.33 (33.04)	50.00 (50.00)	$U = 428$	0.720
Action	64.52 (25.33)	66.60 (31.24)	45.63 (33.10)	47.20 (50.00)	$U = 303.5$	0.029
Outcome	59.00 (22.18)	–	50.00 (27.13)	–	$F = 1.978$	0.165
Reaction	71.10 (23.95)	66.60 (50.00)	53.87 (39.07)	66.60 (87.50)	$U = 356.5$	0.155

Note: The data are presented as mean/standard deviation, in the case of a normal distribution and/or as interquartilic average/range (IQR) in the opposite case.

of the recall of the different categories included in the grammar of the story, the children with ADHD recalled fewer propositions in the different categories. However, the differences are only significant for the proportion of recalled story propositions included in the categories of Introduction, U (58) = 319, $p < 0.046$, Event, U (58) = 297, $p < 0.018$ and Action, U (58) = 303.5, $p < 0.029$, in favor of the control group.

Figure 1 presents the percentages of the recalled propositions in each of the categories for the two groups of subjects who participated in our study. This representation allows us to more clearly observe the differences that are produced between the two in recalling the story categories.

The data indicate that the order of the categories from the best recalled to the worst in the control group were: Event, Reaction, Action, Introduction, Outcome, and Internal Response. For the ADHD group, the order of the categories was the following: Event, Reaction, Outcome, Action, Introduction, and Internal Response. As can be observed, the hierarchy of the categories is very similar in both groups, although it can be pointed out that the subjects with ADHD have better recall of the propositions belonging to the outcome category than to the categories of Action and Introduction.

Results of the text composition measures. The results obtained from comparing the two groups of children, with and without ADHD, on their writing reveal that there was no difference in the time it took them to write the essay, $F(1,58) = 3.32$, $p < 0.074$ (see Table 3).

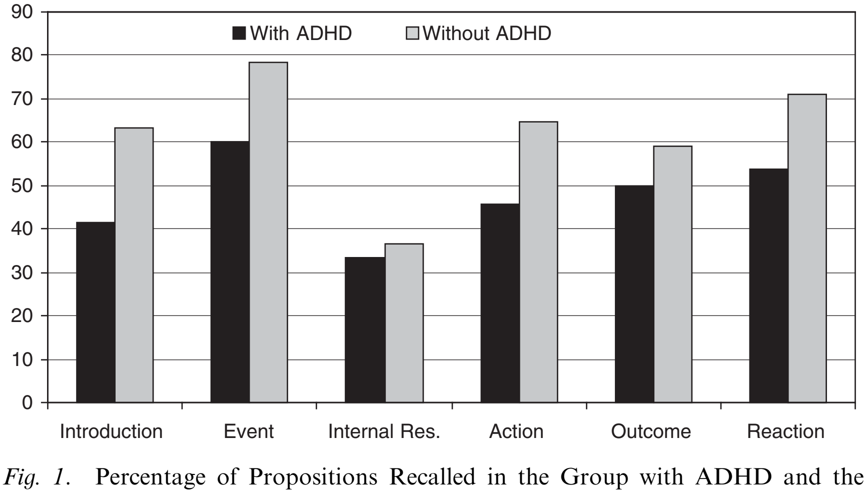

Fig. 1. Percentage of Propositions Recalled in the Group with ADHD and the Group without ADHD in each of the Categories of the Story.

Table 3. Comparison of Subjects with ADHD and Normal Subjects on Variables Related to Text Composition.

	Without ADHA ($N = 30$)		With ADHD ($N = 30$)			
	Mean (DT)	Average (IQR)	Mean (DT)	Average (IQR)	Statistical F/ U	Bilat. Signif.
Time	277.23 (106.10)	–	330.00 (117.94)	–	$F = 3.32$	0.074
Number of words	51.83 (22.72)	–	42.56 (20.30)	–	$F = 2.77$	0.101
Number of sentences	7.56 (3.69)	–	5.30 (2.58)	–	$F = 7.58$	0.008
Mean length sentences	7.59 (3.28)	–	8.44 (2.45)	–	$F = 1.28$	0.263
Syntactic errors	1.70 (2.48)	0.00 (3.00)	3.07 (3.19)	2.00 (4.75)	$U = 314.5$	0.038
Expressive content	10.23 (6.26)	8.50 (6.50)	9.07 (5.75)	8.00 (9.25)	$U = 404.50$	0.500
Type-token ratio (TTR)	0.86 (0.07)	–	0.72 (0.12)	–	$F = 31.61$	0.000

Note: The data are presented as mean/standard deviation, in the case of a normal distribution and/or as interquartilic average/range (IQR) in the opposite case.

The results of the evaluation of the content of the essays written by the children indicate that the essays of the two groups of children did not differ with regard to the number of words, $F(1,58) = 2.77, p < 0.101$. On the other hand, although the children with ADHD wrote a significantly inferior number of sentences than the children without ADHD, $F(1,58) = 7.58$, $p < 0.008$, their sentences do not differ as far as average length is concerned, $F(1,58) = 1.28$, $p < 0.263$.

Furthermore, the essays of the children with and without ADHD do not differ with regard to the type of expressive content, $U(58) = 404.50$, $p < 0.500$. However, there are significant differences in the number of syntactic errors they make, $U(58) = 314.5$, $p < 0.038$ and the degree of variability in the lexicon used by both groups, $F(1,58) = 31.61$, $p < 0.000$, to the detriment of the group of children with ADHD.

DISCUSSION

The purpose of this study was to evaluate the performance of children with ADHD on different text comprehension tasks with different cognitive demands and on text composition tasks.

In the first place, our results show that the children with ADHD without word recognition difficulties did not present problems on either literal or inferential comprehension, as the number of correct responses they provided was similar to that of the children without ADHD. In addition, as in other previous studies (e.g., Ghelani, Sidhu, Jain, & Tannock, 2004), the scores on literal and inferential comprehension of the children with ADHD were within the mean.

On the other hand, there was a deficient performance on the tasks of reading comprehension that presented high cognitive demands, such as ordering fragments, retelling stories and writing texts. Specifically, on the fragment ordering task, which is considered a measure of local and global coherence (Scardamalia & Bereiter, 1984), the performance of the groups of students with ADHD was significantly worse with regard to the accuracy with which they carried out the ordering. This result could be either due to difficulties of children with ADHD in comprehending causal relations, to the low cognitive involvement students with ADHD usually have on long tasks, or to both these factors (Lorch et al., 2004). However, in contrast to what was expected, the number of movements the group with ADHD used in carrying out the task was similar to that of the children without ADHD. Reflecting on the results related both to the accuracy and to the number of movements leads us to hypothesize that on this task the movements of the children with ADHD lack the purposefulness shown by the children in the group without ADHD. In the latter case, it could be said that their movements are directed toward bringing coherence to the text they have to organize. Future research will have to find out whether our hypothesis is justified or not.

Regarding the narration of a story, our findings agree with those of previous studies, even though they used different evaluation procedures (e.g., Lorch et al., 2004b; Renz et al., 2003; Lorch et al., 1999). Thus, the children with ADHD present a performance level that is generally lower than that of normal children on the task of recalling a story they had read earlier. Important differences were observed with regard to the way they structured their narrations. In fact, the children with ADHD remembered significantly less information from the story in the categories of Introduction, Event and Action; that is, they omitted more information about the presentation and events of the story, as well as the actions carried out by the protagonists to reach their goal. The subjects with ADHD did not remember as much as the normal subjects about the structure of the protagonist's goal, the essential element of a story, a result that was also obtained by Renz et al. (2003).

With regard to composing a text, just as in the study by Ross et al. (1995), group differences were not found regarding the speed with which they produced the text. Our results also show, as did the findings from the Resta and Eliot (1994) study, that the students with ADHD are less proficient writers. They used fewer sentences and made a greater number of syntactical errors in them, such as the suppression of functional words (e.g., articles, prepositions, conjunctions) and the repeated enumeration of words in their written texts. It was also observed that the texts of the children with ADHD were less lexically diverse, which indicates that there were more repeated words. However, no differences were found in the average lengths of the sentences.

In general terms, the results obtained by the children with ADHD make it obvious that they have special difficulties on the tasks that require organizing and structuring information, because ADHD is associated with significant weaknesses in several key executive functions domains, which could have a negative impact on academic functioning (Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005; Biederman et al., 2004). Unfortunately, these difficulties can have a large impact on the daily lives of the children with ADHD, not only due to the effect they have on the comprehension and composition of texts, but also because they could denote mistakes in social comprehension (Francis, Fine, & Tannock, 2001). In fact, there are studies that show that children with ADHD experience difficulties when interpreting the actions of other children and in seeing the connections between their own actions and their consequences (Milch-Reich, Campbell, Pelham, Connelly, & Geva, 1999).

METHODOLOGICAL CONSIDERATIONS OF THE PRESENT STUDY AND DIRECTIONS FOR FUTURE RESEARCH

In this chapter, we have provided significant information that documents the reading comprehension and written composition problems of children with ADHD. However, our findings must be interpreted with various strengths and limitations in mind with regard to the method and measures used. In the first place, the subjects with ADHD who participated in our study were diagnosed in a Neuropediatric Service according to the criteria of the DSM-IV. The sample was restricted to children with ADHD without word recognition problems and with average intelligence, so that the findings related to performance on text comprehension would not be affected by the subjects' low intelligence or difficulties with lexical access.

Another aspect of the sample worth highlighting was the inclusion of a control group of children without ADHD. This group was matched with the group of children with ADHD on verbal IQ, vocabulary, and word recognition. By adopting this strategy, it was possible to make sure that any differences between the two groups in their performance on reading comprehension tasks were not due to differences in general verbal abilities or their ability to read isolated words. However, in this study the reading rates of the two groups of children being compared was not controlled. Therefore, it is possible that the reading rate might have had some influence on the low performance shown by the children with ADHD on some of the comprehension and written composition measures used. In future studies on the topic, attention should be paid to controlling this variable.

Obviously, if the sample had included another group of students with difficulties in reading comprehension, crucial information would have been obtained about the possible differences in comprehension processes between the children with ADHD and those with RD, which would make it possible to design intervention programs to fit the specific deficits of each group. Moreover, in order to achieve a more in-depth understanding of the difficulties of children with ADHD on comprehension and composition tasks, future research should also study students of different educational levels.

With regard to the measures used in this study, we would also like to point out various positive points. Given that the complexity involved in comprehending a text requires the coordination of different cognitive and meta-cognitive processes, the decision was made to select four measures of comprehension that presented different cognitive demands: literal comprehension, inferential comprehension, fragment ordering, and story recall. The spectrum of tests used includes both classic tests and tests focused on evaluating specific and overall coherence and organizational skills in retelling a story read earlier. Furthermore, special care was taken in adapting the extension of the stories presented in the inferential comprehension, story recall and fragment ordering, tasks to the developmental level of the participants, in order to eliminate the effect of age on performance. Likewise, the stories that included the three tasks mentioned were constructed maintaining the same causal network, in order to make sure that possible differences would not be determined by this factor. On the contrary, only traditional measures were used for evaluating written discourse. Even so, we believe the results obtained are valuable, given that research on the written composition of children with ADHD is still scarce.

One important question open to future research involves comparing the performance of children with ADHD on comprehension tasks with different text genres, narrative versus expository. This type of texts comprises a variety of structures and usually presents unfamiliar content, and it is quite probable that the reading and writing difficulties of children with ADHD are exacerbated when they are faced with tasks requiring an increased cognitive load, as in expository passages.

On the other hand, this study did not use direct measures of executive functioning (i.e., working memory, inhibition, set shifting, lexical fluency) that would have made it possible to determine which executive processes are involved in the comprehension and composition problems of the children with ADHD and the explanatory role they play in these deficits. Furthermore, the selection of the measures used in our study has been guided by a cognitive approach and, consequently, no implications for a motivational or emotional approach can be drawn from the results of our research. The possibility that motivational factors affect the performance of children with ADHD on comprehension and composition tasks deserves further investigation, particularly the comparison of different methods of assessing reading comprehension and composition skills.

The handling of data and the statistical analyses carried out in this research were rigorous. Following basic statistical proposals, the normality of the distribution of the data was shown by applying the Shapiro–Wilks test. Later, other tests were used for comparing the two groups, with ADHD and without ADHD; these tests were either parametric (ANOVA), or nonparametric (Mann–Whitney *U* test).

In conclusion, we would like to bring up one final consideration. We are aware that in order to answer questions with relevant practical repercussions, the contributions made by quantitative research, like the one presented in this chapter, should be enriched by the contributions of qualitative research. A qualitative methodology makes it possible to identify possible factors that influence the performance of students with ADHD on comprehension and written expression tasks in the classroom on a day-to-day basis. Progress toward a greater "contextualization" and ecological validity of the research carried out is necessary, given that there is a large amount evidence that the cognitive difficulties of children with ADHD are largely dependent on the context.

After the methodological considerations described, we would like to finish by emphasizing the practical relevance of our findings for education. Our study makes it clear that the evaluation of the reading comprehension of students with ADHD should include a wide range of comprehension tasks with different cognitive demands, in order to determine which specific problems these children are experiencing. Furthermore, the psycho-educational evaluation of the students with ADHD should also extend to writing, given that they are at risk for having deficits in written composition skills. Proceeding in this way will provide teachers with sound criteria for developing educational interventions that are tailored to specific deficits in reading comprehension and written expression of students with ADHD.

NOTES

1. A more exhaustive analysis of the narrative abilities of children with ADHD can be found in Miranda, García and Soriano (2005).

REFERENCES

Alarcos, L. E. (1995). *Gramática de la lengua Española*. Madrid: Real Academia Española.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.) (Trad.cast. Manual diagnóstico y estadístico de los trastornos mentales. DSM-IV. Barcelona: Masson.1995).

Barry, T. D., Lyman, R. D., & Klinger, L. G. (2002). Academic underachievement and attention-deficit/hyperactivity disorder: The negative impact of symptom severity on school performance. *Journal of School Psychology*, *40*, 259–283.

Biederman, J., Monuteaux, M. C., Doyle, A. E., Seidman, L. J., Wilens, T. E., Ferrero, F., Morgan, Ch. L., & Faraone, S. V. (2004). Impact of executive function deficits and attention-deficit/hyperactivity disorder (ADHD) on academic outcomes in children. *Journal of Consulting and Clinical Psychology*, *72*(5), 757–766.

Brock, S., & Knapp, P. (1996). Reading comprehension abilities of children with attention-deficit/hyperactivity disorder. *Journal of Attention Disorders*, *1*, 173–186.

Calsamiglia, H., & Tusón, A. (1999). *Las cosas del decir. Manual de análisis del discurso*. Madrid: Ariel.

Cervera, C., & Toro, J. (1984). *TALE. Test de Análisis de la Lectoescritura*. Madrid: Visor.

Cherkes-Julkowski, M., Stolzenberg, J., Hatzes, N., & Madaus, J. (1995). Methodological issues in assessing the relationship among ADD, medication effects and reading performance. *Learning Disabilities: A Multidisciplinary Journal*, *6*, 21–30.

Clemente, R. A. (1989). Medida del desarrollo morfosintáctico. Los problemas de la medición y utilización de la LME. *Anuario de Psicología*, *42*, 103–115.

Clemente, R. A. (1995). *Desarrollo del lenguaje*. Barcelona: Octaedro.

Douglas, V. I. (2005). Cognitive deficits in children with attention deficit hyperactivity disorder: A long-term follow-up. *Canadian Psychology*, *46*, 23–31.

Dubey, D. R., & O'Leary, S. (1975). Increasing reading comprehension of two hyperactive children: Preliminary investigation. *Perceptual and Motor Skills*, *41*, 691–694.

Faraone, S. V., Bierderman, J., Monuteaux, M. C., Doyle, A. E., & Seidman, L. J. (2001). A psychometric measure of learning disability predicts educational failure four years later in boys with attention-deficit /hyperactivity disorder. *Journal of Attention Disorders*, *4*, 220–230.

Francis, S., Fine, J., & Tannock, R. (2001). Methylphenidate selectively improves story retelling in children with attention deficit hyperactivity disorder. *Journal of Child and Adolescent Psychopharmacology*, *11*, 217–228.

Gárate, M. (1994). *La comprensión de cuentos en los niños. Un enfoque cognitivo y sociocultural.* Madrid: Siglo XXI.

Ghelani, K., Sidhu, R., Jain, U., & Tannock, R. (2004). Reading comprehension and reading related abilities in adolescents with reading disabilities and attention-deficit/hyperactivity disorder. *Dyslexia*, *10*, 364–384.

Lorch, E. P., Diener, M. B., Sanchez, R. P., Milich, R., Welsh, R., & van der Broek, P. (1999). The effects of story structure on the recall of children with attention-deficit hyperactivity disorder. *Journal of Educational Psychology*, *91*, 273–283.

Lorch, E. P., Eastham, D., van der Broek, P., Milich, R., Lemberger, C. C., Sanchez, R. P., & Welsh, R. (2004). Difficulties in comprehending causal relations among children with ADHD: The role of cognitive engagement. *Journal of Abnormal Psychology*, *113*, 56–63.

Lorch, E. P., O'Neil, K., Berthiaume, K. S., Milich, R., Eastham, D., & Brooks, T. (2004). Story comprehension and the impact of studying on recall in children with attention deficit hyperactivity disorder. *Journal of Clinical Child and Adolescent Psychology*, *33*, 506–515.

Mayes, S. D., Calhoun, S. L., & Crowell, E. W. (2000). Learning disabilities and ADHD: Overlapping spectrum disorder. *Journal of Learning Disabilities*, *33*, 417–424.

Milch-Reich, S., Campbell, S. B., Pelham, W. E., Connelly, L. M., & Geva, D. (1999). Developmental and individual differences in children's on-line representations of dynamic social events. *Child Development*, *70*, 413–431.

Miranda, A., García, R., & Soriano, M. (2005). Habilidad narrativa de los niños con trastorno por déficit de atención con hiperactividad. *Psicothema*, *17*(2), 227–232.

O'Neill, M. E., & Douglas, V. I. (1991). Study strategies and story recall in attention deficit disorder and reading disability. *Journal of Abnormal Child Psychology*, *19*, 671–692.

Paniagua, G. (1983). El recuerdo de cuentos en niños preescolares. *Infancia y aprendizaje*, *22*, 47–56.

Purvis, K. L., & Tannock, R. (1997). Language abilities in children with attention-deficit hyperactivity disorder, reading disabilities and normal controls. *Journal of Abnormal Child Psychology*, *25*, 133–144.

Renz, K., Lorch, E. P., Millich, R., Lemberger, C., Bodner, A., & Welsh, R. (2003). On-line story representation in boys with attention-deficit hyperactivity disorder. *Journal of Abnormal Child Psychology*, *31*, 93–104.

Resta, S. P., & Eliot, J. (1994). Written expression in boys with attention deficit disorder. *Perceptual and Motor Skills*, *79*, 1131–1138.

Ross, P. A., Poidevant, J. M., & Miner, C. U. (1995). Curriculum-based assessment of writing fluency in children with attention-deficit hyperactivity disorder and normal children. *Reading and Writing Quarterly: Overcoming Learning Difficulties*, *11*, 201–208.

Samuelson, S., Lundberg, I., & Herkner, B. (2004). ADHD and reading disability in male adults: Is there a connection? *Journal of Learning Disabilities*, *37*, 155–168.

Scardamalia, M., & Bereiter, C. (1984). Development of strategies in text processing. In: H. Mandl, N. L. Stein & T. Trabasso (Eds), *Learning and comprehension of text* (pp. 221–254). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Semrud-Clikeman, M., Biederman, J., Sprich-Buckminster, S., Lehamn, B. K., Faraone, S. V., & Norman, D. (1992). Comorbidity between ADHD and learning disability. A review and report in a clinically referred sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*, 439–448.

Serra, M., Serrat, E., Bel, A., & Aparici, M. (2000). *La adquisición del lenguaje*. Madrid: Ariel.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In: R. D. Freedle (Ed.), *Advances in discourse processes: Vol. 2. New directions in discourse processing* (pp. 53–119). Norwood, NJ: Ablex.

Tannock, R., Purvis, K. L., & Schachar, R. J. (1993). Narrative abilities in children with attention-deficit hyperactivity disorder and normal peers. *Journal of Abnormal Child Psychology*, *21*, 103–117.

Webster, R. E., Hall, C. W., Brown, M. B., & Bolen, L. M. (1996). Memory modality differences in children with attention deficit hyperactivity disorder with and without learning disabilities. *Psychology in the Schools*, *33*, 193–201.

Wechsler, D. (1993). *WISC-R. Escala de Inteligencia para Niños Revisada [Wechsler Intelligence Scale for Children]*. Madrid: TEA.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry*, *57*, 1336–1346.

Zentall, S. S. (1988). Production deficiencies in elicited language but not in the spontaneous verbalization of hyperactive children. *Journal of Abnormal Child Psychology*, *16*, 657–673.

APPENDIX 1. STORIES USED IN CHILDREN'S RECALL

Story 1: María and her Duck

Introduction

1. María was a girl
2. who lived on a farm
3. near a river
4. *where many children played.*

Event

5. One day a boat reached the river
6. María went with her friends
7. and saw how a strange bird
8. talked to a sailor on the deck of a boat.

Internal response

9. María was surprised
10. and thought she could teach her duck to say a few words.

Action

11. The next day she got her duck,
12. went to the river near her farm,

13. put the duck on a rock
14. and stood in front of him and asked him to speak.

Outcome
15. *The duck sat looking at María*
16. *moved his head,*
17. jumped off the rock,
18. and went to play with a girl duck.

Reaction
19. That day María realized that ducks do not talk
20. *and that ducks have more fun swimming in ponds and rivers.*

Story 2: The Boy and the Genie

Introduction
1. Once there was a boy
2. who was walking along a path

Event
3. and he met a strange flying box.
4. The boy caught the box

Internal response
5. *thinking he could use it to keep things in*

Action
6. He put it under his arm

Outcome
7. and as he put it there, he heard some mysterious words coming from the box:
8. "you carried me and I am going to help you".

Internal response
9. Scared by these words that came out of the box, the boy

Action
10. opened it very quickly,

Outcome

11. and inside he saw a genie that was brighter than the sun.

Reaction

12. From that day on, the genie has stayed with the boy.
13. *If the boy wanted to spend the day playing, the genie did his homework,*
14. *he told him the questions that would be on the test*
15. *and where he could find the best toys to have fun with.*

Event

16. One day, the genie gave the boy a marble the color of the sun that had magic powers
17. and he told him that with that marble he could help to free a playful puppy
18. that was held prisoner by a witch in a cave.

Action

19. The next day, the boy and the genie traveled to the entrance of the cave.
20. The genie hid inside a vase
21. and started to bark like a dog: "bow, wow".

Outcome

22. The witch came out to see what was going on.

Action

23. The boy took the opportunity and went inside the cave,
24. got the playful puppy out,
25. threw the marble at the head of the witch

Outcome

26. who started to stumble and fell down dead.

Reaction

27. The playful puppy was as happy as could be
28. and they became friends.

CONDUCTING INTEGRATIVE REVIEWS OF SPECIAL EDUCATION RESEARCH: OVERVIEW AND CASE STUDY

Elizabeth A. Edgemon, Andrew L. Wiley, Brian R. Jablonski and John W. Lloyd

ABSTRACT

Integrative reviews are an important method for understanding research in the field of special education. Reviews can help practitioners decide what methods to use in the classroom, researchers clarify directions for new research, and policymakers guide education improvement programs. We discuss the steps for conducting an integrative review, illustrating the process with a case study of an integrative review of large-scale testing accommodations for students with disabilities.

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 257–284

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19010-2

CONDUCTING INTEGRATIVE REVIEWS OF SPECIAL EDUCATION RESEARCH: OVERVIEW AND CASE STUDY

Talk of what "research says" or "research proves" is frequent in special education and other fields. People use the phrase "research says" to garner credibility for whatever claim follows. Usually, when people say "research says," they follow it with a statement about improving schools or students' outcomes: "Research says schools that are more x are better learning environments" or "research says that children learn better when ___." However, making sense of everything special education research is *purported* to say almost certainly will be an exercise in inconsistency, contradiction, imprecision, and overgeneralization.

If special education practice is to improve, deciphering what to infer from reliable research data is a necessary (Carnine, 1997; Kavale, 2001b; Vaughn & Dammann, 2001) if not sufficient (Brigham, Gustashaw, Wiley, & Brigham, 2004; Landrum, 1997) precondition. Current legislative mandates (i.e., the No Child Left Behind (NCLB) and Individuals With Disabilities Educational Improvement Acts) that expressly encourage use of research-validated classroom methods still leave the field of special education with the problem of determining what methods have and have not been validated, and how various methods compare to one another (Jensen, 2003; Odom et al., 2005).

Objective reviews of the research about special education procedures, methods, and practices help overcome these difficulties. Integrative review techniques synthesize information from many studies (Cooper & Hedges, 1994; Glass, 1976; Lipsey & Wilson, 2001). Importantly, integrative techniques allow us to derive data-based generalizations that are more objective, verifiable, and replicable than mere assertion (Kavale, 2001b; Kavale & Forness, 2000; Lloyd, Pullen, Tankersley, & Lloyd, 2006). Integrative research reviews have been and will be essential to progress in both applied and theoretical aspects of special education.

In this chapter, we outline techniques for carrying out integrative research reviews and illustrate them by showing how they apply to a review accommodations on high-stakes tests. First, we describe the basic elements, advantages, and limitations of integrative literature reviews, as well as a necessarily brief sample of reviews that have been influential in special education research. Next, we present a fundamental sequence of steps for conducting integrative reviews. We describe our own experiences reviewing testing accommodation research (Jablonski, Edgemon, Wiley, & Lloyd,

2005) to illustrate the steps and to highlight some practical problems encountered in research synthesis.

Integrative Reviews – Basic Elements

Integrative review is a form of research in its own right based on questions about the relationship between independent and dependent variables. Studies that report primary data are treated as "subjects," and systematic analysis is employed to reach conclusions that are valid and reliable (Kavale & Forness, 2000). Cooper, Valentine, and Charlton (2000) offer a five-stage model of research synthesis. The stages include problem formulation, data collection, data evaluation, data analysis and interpretation, and public presentation. We cover these stages or steps later in this chapter. Although the similarities to other types of research are numerous, analysis and synthesis of data from *other studies* clearly differentiates integrative review from other activities. Cooper and Hedges define an integrative research review as an "attempt to integrate empirical research for the purpose of creating generalizations" (1994, p. 5).

Reviewers seek to create data-based generalizations for a variety of reasons. Kavale and Forness (2000) note that research synthesis can be used to clarify the parameters of a given phenomenon, to place a phenomenon in context, to make explicit what is only implicit, and to eliminate unessential elements by providing a conceptual whole. Synthesis of intervention research helps separate components that underlie treatment effectiveness, permitting increased efficiency in both implementation and further research and development. Careful synthesis permits us to examine the existence and strength of a cause (e.g., treatment) and effect (e.g., outcome) relationship (Hall, Tickle-Degnen, Rosenthal, & Mosteller, 1994). Additionally, integrative reviews can be used to resolve conflicts, present criticism, identify central issues, and to identify research shortcomings (Cooper, 1988). In addition to their potential for improving practice, integrative research reviews can contribute to the refinement of theory by aggregating "estimates of magnitudes of effects for theorized relations that rarely would or could be tested within one primary study" (Hall et al., 1994, p. 21).

Integrative reviews differ from each other in a variety of other respects. Reviews can be neutral, or they can advocate a theoretical, conceptual, or practical position. The research covered can be exhaustive or selective, and it can be organized from earliest to latest, conceptually, or by the research methods employed. Reviews can be written for other researchers, educators, the general public, or policymakers (Cooper, 1988). Finally, and perhaps most importantly, integrative research reviews can be characterized by the extent to which they do or do not use quantitative-statistical (meta-analytic) techniques to analyze data from primary studies statistically (Kavale, 2001b). *Narrative* reviews report the preponderance and general directionality of the evidence, as understood by the reviewer; although narrative reviews are limited in their exactitude, they are useful when the literature does not lend itself to meta-analysis (Mostert & Kavale, 2001).

The development of *meta-analytic* review techniques represented an improvement over narrative reviews because meta-analysis "made [synthesis] processes public and based them on explicit, shared, statistical assumptions" (Cooper & Hedges, 1994, p. 11). Essentially, meta-analysis involves the calculation of a standardized measure of the effect of an independent variable on a dependent variable. Because the effect size (ES) statistic is standardized, comparisons of effects can be made across interventions, dependent variables, studies, and meta-analyses. Although experts disagree about standards for what represents an important effect, ESs of less than approximately 0.2 are clearly weaker and ESs of approximately 0.6 and greater are considered relatively strong (Lloyd, Forness, & Kavale, 1998). Meta-analytic methods succeed where narrative reviews fall short in terms of reliably and objectively organizing, extracting, and accumulating usable and interpretable findings from large databases (Kavale & Forness, 2000).

Integrative Reviews – Advantages and Limitations

Achieving synthesis in special education research is complicated by the wide variety of research designs, outcome measures, and diagnostic labels used (Cooper et al., 2000), not to mention the intensely emotional nature of disability discourse. Given the complexities of the field, the problem is not just that single studies are unlikely to provide "definitive answers" for teachers and policymakers; the problem is also that answers are likely to be conditional (Kavale & Forness, 2000; Matt & Cook, 1994). Integrative reviews that apply meta-analytic techniques tell us not only what works, but how much, in comparison to what, for whom, and under what conditions things work (Cooper et al., 2000; Kavale, 2001b). Integrative reviews point out important modifiers and limitations of generalizations to researchers, practitioners, and policymakers (Cooper & Hedges, 1994). Including the population of relevant studies (rather than a sample), with post-hoc analysis of interactions between study variables and outcomes, eliminates or reduces bias created by selectively sampling research literature (Kavale, 2001b). In a field vulnerable to the empty promises of pseudoscience and ideology, integrative reviews offer safeguards via explicit and objective renderings of what ‘‘research says’’ (Kavale & Forness, 2000) as well as transparency and accessibility to scrutiny (Lipsey & Wilson, 2001).

Integrative research reviews have several limitations which reviewers, educators, and policymakers should keep in mind. As with primary research, findings from integrative research reviews fall along a continuum of reliability and validity. Threats to the internal validity of integrative review findings may arise from methodological flaws, including how the research problem is formulated, how conceptual distinctions are drawn, and how data are gathered and evaluated (Cooper et al., 2000). Unclear conceptual distinctions and poor problem formulations may lead to ‘‘apples to oranges’’ comparisons (Lipsey & Wilson, 2001). The methodological quality of existing meta-analyses in special education is difficult to determine when reviewers do not include detailed procedures (Mostert, 2001). The methods of integrative research review and meta-analysis are flexible and often complex; each technique requires some degree of judgment and thoughtful decision-making guided by experience, expertise, and the research problem in question (Kavale, 2001b; Lipsey & Wilson, 2001). Also, the reliability of generalizations derived from research syntheses is largely dependent on the reliability of the data from the primary studies reviewed (Matt & Cook, 1994). Good reviews will test the moderating influence of study quality; however, if no high-quality studies can be located and included in the review, generalizations will be suspect. Last, there is no foolproof way to know that the entire population of relevant studies has been included (Kavale, 2001b). For these and other reasons (Cooper & Hedges, 1994), stakeholders should be cautious in interpreting findings from integrative research reviews.

Integrative Research Reviews – Influential Examples from Special Education Research

Since the 1980s, special education researchers have recognized and made use of the unique potential of integrative research review techniques. Not surprisingly, most research syntheses in special education have examined the relative effectiveness of educational interventions for students with learning and behavioral problems. Special education meta-analyses of intervention research have revealed ESs that range from disappointing to extremely encouraging (Lloyd et al., 1998). Strong support has been found, for example, for school-based behavior reduction strategies (Stage & Quiroz, 1997); reading comprehension instruction for students with learning disabilities (LD) (Swanson, 1999; Talbott, Lloyd, & Tankersley, 1994); early intervention (Casto & Mastropieri, 1986); cognitive behavior modification (Robinson, Smith, Miller, & Brownell, 1999); systematic formative evaluation (Fuchs & Fuchs, 1986); mnemonics (Mastropieri & Scruggs, 1989); and direct instruction (White, 1988).

Integrative research review techniques have permitted fine-grained analyses of intervention outcome data. For example, Swanson and Hoskyn (2001) used meta-analysis to pinpoint two instructional components (advanced organization and explicit practice) that account for significant variance in the learning of adolescents with learning disabilities. The observation by Kavale (2001a) that interactions outnumber main effects in special education intervention research has been borne out repeatedly. For instance, interventions for attention-deficit hyperactivity disorder (DuPaul & Eckert, 1997) and disruptive behavior (Stage & Quiroz, 1997) have proven most effective when delivered via special education. Talbott et al. (1994) found reading comprehension interventions to be more effective when implemented by the study author as well as with older children.

Integrative reviews have also exposed the weak empirical bases of numerous popular special education interventions. Small-to-negligible ESs have been found, for example, for social skills training (Forness & Kavale, 1996; Quinn, Kavale, Mathur, Rutherford, & Forness, 1999); modality-based instruction (Kavale & Forness, 1987); special diet for hyperactivity (Kavale & Forness, 1983); and perceptual-motor training (Kavale & Mattson, 1983). Reviews have determined that some widely advocated and implemented practices have almost no or insufficient research to recommend them; e.g., co-teaching (Murawski & Swanson, 2001; Weiss & Brigham, 2000) and functional behavior assessment for students with emotional and behavioral disorders (Nelson, Roberts, & Mathur, 1999; Sasso, Conroy, Stichter, & Fox, 2001). Negative reviews of some interventions have lead to counter-reviews meant to refute the findings of the first (e.g., a meta-analytic defense of learning styles by Dunn, Griggs, Olson, Beasley, & Gorman, 1995). Owing to the ultimately transparent nature of integrative research reviews, claims and counter-claims about effectiveness can nonetheless be judged relative to the quality of the evidence on which they are based, rather than the arbitrary authority of the claimant. Forness (2001) suggested that a larger pattern has emerged from meta-analyses, suggesting that interventions characterized by direct efforts to achieve tangible goals have the biggest effects, whereas indirect efforts to change underlying problems have had minimal benefits.

Integrative research reviews have been used in special education research for purposes other than evaluation of interventions. Other purposes include clarifying central issues, identifying the parameters of phenomena, pointing out gaps in the knowledge base, and trying to resolve conflicts. For example, Cameron and Pierce (1994) conducted a meta-analysis and concluded that the widely held belief that extrinsic rewards diminish intrinsic motivation, except under very specific and avoidable conditions, has little empirical backing. Deci, Koestner, and Ryan (1999) conducted their own meta-analysis and found that extrinsic rewards *do* diminish intrinsic motivation; however, Deci et al. *took a theoretical position* and selectively included studies in which participants were rewarded for things they were already motivated to do. Elbaum (2002) used meta-analysis to assess conflicting assumptions about the self-concept of students with learning disabilities. One assumption holds that labeling and segregating students with learning disabilities lowers their self-concept; the other suggests that self-concept develops by comparing oneself to peers, suggesting that being taught with other students with learning disabilities might boost self-concept. Finally, integrative reviews have been used to quantify the academic deficits of students with emotional and behavioral disorders (Reid, Gonzalez, Nordness, Trout, & Epstein, 2005), the social skill deficits of students with learning disabilities (Kavale & Forness, 1996; Swanson & Malone, 1992), and compare characteristics of students identified as having different categories of disability (Sabornie, Culuinan, Osborne, & Brock, 2005) and to describe characteristics of students who do not benefit from early reading intervention (Al-Otaiba & Fuchs, 2002). These are just a few examples of the usefulness of research syntheses beyond appraisals of the relative value of various interventions.

BEGINNING THE REVIEW

As with any study, the strength of an integrative research review depends in part on the focus of the study and on a well-defined and narrow research question. A research question that is too narrow will not yield enough studies for the review to be summative and draw conclusions, whereas a question that is too broad could take years, attempting to answer so many questions that the reviewers find it difficult to answer any questions.

Formulate a Problem

Reviews begin with clarification of the problem or phenomenon of interest (Lipsey & Wilson, 2001). Previous reviews, influential position papers, and personal experience and knowledge may provide direction for where to begin. Integrated research reviews can be used to describe relationships between independent and dependent variables, refine theory, resolve conflicts, and identify central issues (Cooper & Hedges, 1994).

Case study. As special educators, we were interested in accommodations on large-scale tests for students with disabilities. Specifically, we wanted to determine what accommodations students with disabilities need and should receive on large-scale tests, under what conditions accommodations are appropriate, what the purpose of accommodations is, and what the effects of large-scale testing accommodations are. Thus, we identified a topic and defined the purposes of our review by asking relevant questions (Kavale & Forness, 2000).

Review secondary sources. Once a topic has been identified, the researcher explores the literature on the subject. If similar reviews have been done recently, the researcher must determine what important questions remain unanswered. Older reviews can help direct and narrow the focus of the proposed review. Additionally, because the concept behind an integrative review is to read and synthesize the works of other researchers, one must make certain that adequate primary research exists (Mertens, 2005). There are no set guidelines for how much is enough, so researchers will have to ask themselves what would be accomplished by compiling information from a small set of studies. Beginning with a broad idea, running preliminary searches, and reading some of the literature helps inform a well-defined research topic.

Refine the question. After reading a sample of studies, the researcher is able to narrow the scope of his study. Though it may seem ideal to keep the focus broad and include all literature related, even vaguely, to the topic, a more focused review will result in clearer conclusions (Cooper & Hedges, 1994). Boundaries for the review must be established with clear research questions and hypothesis statements.

Case study. To obtain a broad overview of the topic of accommodations in our review (Jablonski et al., 2005), we first conducted hand searches of prominent special education journals. Additionally, we conducted a search of the National Center on Educational Outcomes (NCEO) Web site, because it contains relevant reports on the use of accommodations on large-scale tests. Reading position papers, previous reviews, and a sample of primary research added to our understanding of current problems in providing and researching accommodations.

In our accommodations review, we limited the scope of our review after surveying the relevant literature. We decided to include only empirical studies of students with disabilities in grades K-12, their use of accommodations, and their performance on large-scale tests. These rules eliminated studies of students in college or in distance education programs, those that did not identify students with disabilities, studies concerned only with classroom assessments, surveys of accommodation policies, studies that explored test-taking skills, and those that were themselves literature reviews. We refined our research question: Whether the effects of various large-scale testing accommodations differ depending on the characteristics of the individual being tested (i.e., type of disability, severity of disability, age) and the nature of the test (i.e., subject area, type of test).

Literature Search and Retrieval

Develop a search strategy. It is important to have a system for searching the literature, especially when more than one person is involved in the review. Communication among reviewers facilitates the review's progress, while making sure work is not duplicated. One way to do this is to have a master database for articles that have been considered. As a group, identify primary resource journals for hand searches and methods of obtaining a list of studies to begin the review with, such as Web-based databases. Once the search gets underway, as a researcher reads an article he should also read the references and determine if they are relevant to the review, adding them to the master "obtain" list if they are, and to a "not relevant" list if they are not relevant (Mertens, 2005). Finally, researchers should identify key researchers in the field and contact them to ascertain if research is currently underway.

Conduct searches and select titles. It is easy to conduct a review if you physically have a copy of the article. As an author writes, he may find that he wants to refer to the article, or that something in the coding system is confusing. Having the article will make it easier to address these issues without having to make another trip to the library. Because many reviews limit themselves to peer-reviewed journals, sources are frequently obtained by making copies of the relevant articles from published journals (or retrieving electronic copies of more recent articles). Articles from the ERIC database are also often available online or on microfiche at the library (Mertens, 2005).

Search for fugitive literature. Capturing the full corpus of literature on a topic requires pursuit of less readily available studies. When studies are not published in a journal that is easily accessible, it may take some work and time to track down a copy. Librarians can often help find studies that are in obscure journals, sometimes for a fee. Researchers should also look at newer journals that are not yet indexed as well as unpublished sources, including technical reports, theses and dissertations, papers presented at conferences and meetings, Web-based publications, and "desk drawer" studies with data that were never disseminated.

Continue to use ancestral searches. With each new study, reviewers should check the references. Eventually the reference list becomes familiar, as the reviewer reaches the extent of the previous research. When the reference lists of newly found documents do not include previously unknown sources, one is nearing exhaustion of the literature base, if not of personal energy.

First read of research. As researchers read their population of studies, it will become necessary again to focus the research question. Though some studies will clearly pertain to your stated goal and others will clearly not, there will be an ambiguous area which the entire research team needs to define as issues arise.

Case study. For our review (Jablonski et al., 2005), we recognized that some journals of interest to us are not in the PsychInfo database, because they are more educational in nature, and that ERIC would include those journal citations. However, ERIC is missing citations for a period of 2 years, so we conducted hand searches of the following journals for that time period: *Assessment for Effective Intervention*, *Exceptional Children*, *Journal of Educational Measurement*, *Journal of Educational Research*, *Journal of Learning Disabilities*, *Journal of Special Education*, *Learning Disabilities Research & Practice*, *Remedial and Special Education*. As we read, we simultaneously conducted ancestral searches. We did not have a central clearinghouse for the studies we had obtained or read until we were several weeks into the process, but once that was in place we were much more efficient with our reviewing and reading time. As we read we noticed that some cited studies were unpublished. We searched for them through Web sites as well as by contacting the authors. Some authors were very accommodating and sent us the referenced work; others were difficult to locate. We were never able to make contact with one author, despite repeated e-mails and phone calls, and thus had to eliminate those missing studies. In order to be sure we included all in press or unpublished studies, we put a notice through a special education professional listserve soliciting these studies. In this way we received three additional studies.

As we read through the research for the first time, we more clearly delineated our research interest. We excluded most GED studies, because they involved adults, but included SAT and ACT studies because they included high school students. We decided to accept students who were labeled "reading disabled," a category not recognized in the Individuals with Disabilities Education Act (IDEA), but only if the study included criteria for the label. Communicating these questions and resolving them, as they developed ensured that all of the authors were eliminating and keeping studies based on the same criteria.

CODING

Coding studies is an ongoing process that begins when the reviewers initially survey literature on the topic. What the reviewers decide to code will depend on the research question. Before they begin, the reviewers will have some ideas about what variables are most important; however, after reading several studies they may find that other variables are important to the review.

Lipsey and Wilson (2001) suggest that coding has two parts. The first part contains the codes that describe the study's characteristics (study descriptors). The second part contains the codes that describe the empirical findings of the study (study results). Study descriptors include information about the participants, methods, and treatments (Lipsey, 1994) as well as publication information (Lipsey & Wilson) and judgments of study quality (Wortman, 1994).

Lipsey and Wilson (2001) encourage reviewers to code as much detail as they can find in the studies reviewed. Beginning with a more detailed coding scheme will save time spent later going back through studies to code variables not considered at the start. For example, ES is usually computed by dividing the difference in treatment and control group means by the pooled standard deviation. The pooled standard deviation can easily be computed using the group sizes and standard deviations. If the reviewers begin by coding only for the total number of participants, they will not have the information needed to compute ESs for studies that do not provide a pooled standard deviation.

Stock (1994) urges some caution when adding items to the coding scheme. Each item added increases the amount of time necessary to code the literature, potentially, with no added benefit. Creating a highly detailed coding scheme may also add error to the coding process. Just as complexity in behavioral observation systems reduces inter-observer agreement (Dorsey, Nelson, & Hayes, 1986), complex coding schemes will have more room for error in the coding and the analysis. Additionally, the added variables may add noise to the data and increase the probability of reporting chance relationships (Stock). It is the reviewers' job to determine early in the review process what items will be important for coding based on an understanding of the domain under study. With this in mind, we discuss some areas typically coded in review studies.

Identifying Variables of Interest

Most reviewers will code studies for author, publication, and source as a means to identify studies with which they are working. This information also has the potential to identify important issues. In our review of accommodations for large-scale assessments, we found that the majority of studies were published after 1997. It was in 1997 that the amendments to the IDEA made participation in large-scale assessments mandatory for students with disabilities. In addition to the other findings of our study, simple publication information showed that the increase in research on testing accommodations appears to illustrate legislation's effects on research.

Of course, it is essential to code for other features of the research itself. What factors (pupil age? gender? type of assessment?) might affect outcomes? Reviewers need to identify potential mediating and moderating variables. Theoretical treatments of the topic being reviewed can provide guidance about mediators and moderators. If, for example, a critic of an approach contends that a teaching procedure may work for learners with lower skills, but not for those with greater skills, reviewers can turn this point into a testable hypothesis by coding for student skill level.

Coding for Quality

Wortman (1994) equates quality with relevance and acceptability. Relevance is described as a combination of construct validity and external validity. Internal validity and statistical conclusion validity describe acceptability. Wortman's quality descriptors are the means by which we judge studies' attempts to limit possible alternative explanations. Studies that use control groups with random assignment can usually account for participants' natural change in the dependent variable, whereas it becomes more difficult to make the claim that the treatment caused the effect without the control group.

Validity of studies. Campbell and Stanley (1966) describe 12 factors that jeopardize internal and external validity in experimental and quasi-experimental research. Cook and Campbell (1979) expanded this list to 33 different threats. Chalmers et al. (1981) suggest an alternative list of threats that focus more on the construct validity and statistical conclusion validity. The use of control groups with random assignment eliminates most threats to external validity and construct validity. Various issues in social science research, however, make it difficult (or practically impossible) to use control groups or random assignment. Seethaler and Fuchs (2005) reviewed five journals that frequently published studies related to special education; only 4.2% of these studies used random assignment. Use of control groups and random assignments can be unethical and illegal. We do not expect to see a true experiment on the efficacy of special education placement. The random assignment of students meeting the criteria for special education eligibility to a control group (no special education services) would be illegal under current law and unethical for not providing an appropriate education to vulnerable students.

The National Reading Panel adapted this strategy in its best-evidence analysis of early reading instruction by reviewing only research published in peer-reviewed journals (National Institute of Child Health and Development, 2000). Researchers employ different strategies to control for the quality of the studies examined in reviews. To establish a higher level of quality, Dush, Hurt, and Schroeder (1989) limited their review to studies that had a control group, random or random-stratified assignment to groups, and used standardized or objective measures of the dependent variable.

Coding for the type of study or publication can also lead to analysis that is more precise. Xin and Jitendra (1999) examined publication bias in their meta-analysis of teaching students with LD different problem-solving strategies. They found that unpublished studies had a larger ES that was statistically different from that of the published studies.

Case example. In our review of the use and effects of accommodations for students with disabilities on large-scale assessments (Jablonski et al., 2005), we were reluctant to exclude any studies because of the limited amount of research on the topic. We included both experimental studies and correlation studies with the understanding that correlation studies would not allow us to describe causal relationships; to separate the effects of these two types, however, we coded for type in our database. We also noted that many studies were represented in different forms of literature. The authors of one study had presented the findings at three different meetings before publishing the results in a peer-reviewed journal. Several studies were dissertations that the author subsequently published in a journal. Many other papers had been reported directly to government agencies or other organizations, not published formally.

To avoid describing the same study several times in our review or giving one study more weight, because the authors had actively promoted it in different meetings, we decided a study could only be included once. We defined a study as the quest to answer a specific question. Thus, if two reports used the same population, and took place simultaneously, but looked at different questions, we would include the study. For example, a study by Fuchs (2000) looked at the effects of accommodations on students' test performance in reading using differential item functioning. This study appeared to use the same data Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch (2000) used to answer questions about teachers' choice of accommodations for their students. Because the reports asked different questions, both were included even though they used the same population.

Because we had multiple reports of the same data, we developed a ranking system for inclusion. Studies published in peer-reviewed journals would trump all other forms of publication. Our hierarchy of preference continued with technical reports, dissertations, and unpublished papers in this order.

Description of Participants and Setting

Substantive issues. Lipsey and Wilson (2001) refer to information that deals with the characteristics of the study as "substantive issues." Knowing the characteristics of the different populations allow the reviewers to look for potential mediating variables. In addition to the participants' demographic information (e.g., age, gender, ethnicity, etc.), Lipsey and Wilson suggest reviewers code for personal characteristics (e.g., scores on standardized tests for cognitive ability or personality traits), diagnostic categorization (e.g., learning disabled, depressed, etc.), and setting (e.g., general education classroom, special day school, etc.). Stage and Quiroz (1997) demonstrate the importance of setting in a meta-analysis of cognitive behavior modification. They found that the behavioral intervention is more effective as the settings become more restrictive.

Reviewers should code for information that describes the treatment or intervention. This group of variables includes descriptions of the independent variables (treatments or interventions), such as "general description and type, theoretical orientation, levels represented (e.g., dose, intensity, duration, etc.), organizational characteristics (e.g., age, size, administrative structure), mode of treatment delivery, characteristics of intervention staff or personnel" (Lipsey &Wilson, 2001, p. 85).

Each subject area reviewed will have variables the reviewers will want to code, but are not found in reviews of other topics, especially for variables describing the independent variables. For example, Elbaum (2002) in a study of self-concept coded for different domains of self-concept (e.g., academic, physical, social, etc.). Kroesbergen and Van Luit (2003) in a meta-analysis looking at mathematics problem-solving instruction, not only coded for the type of mathematical instruction (e.g., direct instruction, self-instruction, etc.), they also coded for whether the study used computer-aided instruction and peer tutoring. Later analysis showed that interventions including computer-aided instruction had lower ESs.

Case study. In Jablonski et al. (2005), we originally coded for population demographics including grade level (elementary, middle, and high school) and disability (e.g., learning disabled, hearing impaired, etc.). As we read the research we found that studies included students with different disabilities, but frequently did not differentiate the results by disability. For this reason, we created a code for whether studies differentiated results by disability. Our coding included performance area of the test (e.g., reading, mathematics, history, etc.), whether the study differentiated results by performance area, the assessment instrument, and the accommodations used. We coded for 29 different assessment instruments including assessments used nationally (e.g., the Iowa Tests of Basic Skills, the SAT, etc.) and individual state assessments (e.g., the Kentucky Instructional Results Information System). We grouped accommodations in five areas based on the initial reading of research in the area of accommodations (presentation, timing and scheduling, setting, response, and aides). These groups of accommodations in each area included 5–21 different accommodations.

Coding the Methods

Lipsey and Wilson (2001) suggest that reviewers code "all those methodological and procedural variables that can be coded from the studies and that could conceivably affect study results" (p. 84). They recommend that reviewers code for design of the study, the nature of the control condition (e.g., placebo, wait list, no treatment, etc.), data collection procedures, and data analysis procedures.

Meta-analysis and other reviews are useful for making comparisons based on criteria as well as experimental conditions. In a review of writing strategy instruction, Graham and Harris (2003) coded for who taught the strategy as well as characteristics of the students and the classroom. D. Fuchs, Fuchs, Mathes, and Lipsey (2000) conducted a meta-analysis of studies looking at the differences between students with reading difficulties identified as LD compared with similar students not identified as LD. Because D. Fuchs, Fuchs, McMaster, and Al-Otaiba (2000) compared population groups, demographic factors (e.g., IQ and socio-economic status, SES) became important when looking for independent variables that could explain why one group was identified as LD and another was not.

Empirical Findings

The empirical findings of a study are reports of how an independent variable affected a dependent variable. Integrative reviewers describe the independent variable with the methods coding. The dependent variable will frequently involve numerical variables specific to each study, e.g., means, standard deviations, and sample sizes (Lipsey & Wilson, 2001).

Statistical variables. Reviewers interested in meta-analysis should code for a wide variety of the different variables used for computing the ES, as different studies will report different statistical results. These variables include, but are not limited to, control-group and treatment-group means, standard deviations, and number of participants in each group. For studies that do not have means and standard deviations, the coder will need to rely on *t*-statistics, *F*-tests, or correlation coefficients.

Case study. In our study of accommodations usage (Jablonski et al., 2005), we used a narrative description of the findings. This was, in large part, due to the different natures of the studies we reviewed. Some studies looked at the effects of accommodations on students, but other studies looked at the effects on the test questions. One group of studies described the characteristics of students who used the accommodations. We felt that we would best capture this variability in results by a narrative description.

Systematic Coding

Defining the codes. Stock (1994) recommends that reviewers develop a numerical coding scheme for ease of data analysis using computers. The reviewers should create a codebook early in the review process. This codebook will contain a set of numerical values for each variable. Codes should be exhaustive and mutually exclusive. Many variables will need a code for non-reported data. The reviewers should also create a form or sheet to record coded variables. This can be done with an electronic database or on paper. The paper copy will allow the coder to attach the completed form to the study. This allows for greater ease when the reviewers want to recheck codes in a study. The electronic version will save the time transferring data from paper into a database, spreadsheet, or word processor.

Typically, reviewers will use a combination of dichotomous, categorical, and continuous variables. D. Fuchs et al. (2000) used dichotomous variables that required the coder to answer a yes/no question such as, ''Do LD and LA subjects receive reading instruction in the same setting?'' (p. 88). They also used categorical variables for factors such as grade level and SES, and continuous variables for factors like IQ scores.

Some reviewers will create a narrative coding system; these studies will generally not be meta-analytic. D. Fuchs, Fuchs, McMaster, and Al-Otaiba (2003) coded studies looking at students' response to intervention. They included descriptive phrases for demographic, treatment, and outcome variables. Even if reviewers use an explicit codebook such as used by D. Fuchs et al. (2000), they probably should include at least one narrative variable into which they can record peculiarities of studies.

Starting with a detailed codebook, coders should conduct a test run with five to ten studies (Stock, 1994). This will allow the investigator to note potential difficulties with the codebook and coders to develop an understanding of the definitions in the codebook. This process may also call attention to needed revisions to the codebook and form.

Case study. After developing a list of variables we considered important, we searched for possible values for these variables. To code the accommodations, we conducted a survey of large-scale accommodations lists in five states (Arizona, California, Massachusetts, Missouri, and Virginia). This survey helped us to frame our coding scheme for accommodations into the five general categories mentioned previously, as well as providing an extensive list of accommodations. Communication between coders was extremely important – when we came across something that needed a new code (e.g., another state assessment or an accommodation we had not previously included) we created one and immediately e-mailed that code to the other coders.

As noted previously, we also had three narrative codes; one of these, population, overlapped with other coded variables. Although we coded for the number of participants in the study, their grade level, and their disabilities, the narrative variable allowed us to add other descriptors that we did not consider integral to our question (e.g., setting, gender, and ethnicity) and descriptors that were included in only a few studies (e.g., SES).

Evaluating Coding Decisions

Errors in coding. Errors can come from different sources. Orwin (1994) notes that coder error can result from studies that do not clearly state important information, ambiguous or complex coding factors, bias of the coder, and mistakes by the coder. Integrative reviewers can reduce error by providing coders with appropriate training, pilot testing the codebook and forms, and frequent communication between the reviewers and coders about difficulties.

To reduce errors it is important to assess the extent different coders record the same data for studies. Reviewers can evaluate inter-coder agreement in different ways. A listwise method would count all variables for each study as one agreement; that is if all coders agree on all variables, it is one agreement. Reviewers can also count agreements by item; each variable that every coder has scored the same is counted as one agreement. Reviewers may want to create a hierarchical system (see Orwin, 1994) to recognize that some coding choices are dependent on others.

Inter-scorer agreement is frequently scored as a percent of agreement; the number of agreements achieved divided by all possible agreements. This may result in an inflated sense of agreement. A more conservative approach is Cohen's K (Cohen, 1960). We recommend that reviewers who wish a more thorough discussion of inter-scorer agreement consult Hartman (1977, 1982).

Case study. In our study (Jablonski et al., 2005), the first two authors coded one article together before coding any independently and discussed disagreement; the third author reviewed coding with one of the other two, then coded independently. We compared coding for an initial set of 10 studies. Inter-coder agreement was 87% for all three coders on this set of studies using item-by-item agreement methods. Disagreements fell into two categories: an omission by a coder and differences in interpretation of codes between coders. Disagreements were frequently the result of many data points within one study. For example, one study reported data on 17 different accommodations, which different coders grouped or coded differently. We discussed disagreements prior to resuming coding, and reached agreement on the appropriate coding for the specific instance and subsequent occurrences.

Handling Missing Data

Reviewers will likely encounter three different types of missing data: missing studies, missing demographics, and missing ESs (Pigott, 1994). One reason for using only studies published in peer-reviewed journals as a criterion for inclusion is insuring access (in addition to insuring study quality). As noted previously, studies will often contain vague or imprecise reporting of the results. As Orwin (1994) notes, the reviewer can communicate directly with the author of the study for clarification.

Statistical implications. Before discussing what to do about missing data, it is helpful to have some idea of the types of missing data. We can classify missing data into three areas: missing completely at random (MCAR), missing at random (MAR), and missing due to the variables under consideration. MCAR describes data that are randomly missing and not related to any aspect of the studies or review. MAR describes data that are missing for reasons related to the independent variable, not the dependent variable. Reviewers can ignore MCAR data. MAR data can also be ignored, but with caution. Data that are missing due to the dependent variable should not be ignored. One of the concerns of integrative research is that published studies tend not to include studies with weak or negative results. Integrative reviewers should make all attempts to locate studies that are unpublished, if only to note whether the unpublished data have a different ES (Xin & Jitendra, 1999). Sherman (2000) provides a more detailed description of MCAR and MAR.

Statistical programs such as SAS and SPSS will offer the choice of item or listwise deletion for missing data; these are both appropriate for data that is MCAR or MAR. Other programs will conduct imputations for missing data. Typically, imputation involves making estimates for the missing data based on other variables. Imputation has the potential to result in a more accurate synthesis for data that are not MAR, including data that are assumed to be MAR (Davey, Shanahan, & Schafer, 2001).

Case study. As expected, we had difficulty locating studies, because we adopted a liberal criterion about what products to include in our corpus of studies. We had at least one presentation that was of interest, but the author had moved and we were unable to locate him. One author of multiple studies did not respond to repeated requests for help locating studies. The NCEO, however, was helpful in passing on copies of several studies. We did not make corrections for missing data.

Managing the Database

Reviewers will want to maintain the coded studies in a format that provides easy access. The number of studies and the number of factors coded will influence the type of database. It is possible to use a table in a word-processed document to manage a few factors with a high-narrative content. Spreadsheet programs offer the advantage of sorting data. The analyst can also import spreadsheets directly into statistical programs such as SPSS or SAS. Database programs are better for managing larger files. They can be translated into a spreadsheet and then to a statistical program. Reviewers can use database programs to create a form for data entry making the coders' work easier. Many database programs allow the reviewer to create relational databases. One of the main advantages of relational database programs is the ability to code characteristics of the studies in one database and study results in another, then link the two databases. This promotes efficiency when coding multiple effects from the same study. The coder will record the characteristics of the study only once, but all that information will be linked to each of the effects from the study.

Case study. In Jablonski et al. (2005), we maintained our data in a spreadsheet. This had the advantage of familiarity with each of the coders; no one had to learn to use a new program. It proved easy to add in work from each coder and the sort function was useful in finding studies that met particular criteria. There were also limitations. For example, it was difficult to change preset limits on cell sizes.

DESCRIBING THE COMBINED STUDIES

Once data have been collected, reviewers need to analyze the findings and synthesize the results in a meaningful way. Though pure literature reviews of non-quantitative data include summative paragraphs for each study, integrative reviews including quantitative studies need to objectively combine data from studies that have procedural differences. This is done using ESs.

Effect sizes from group designs. Reviewers who wish to do a meta-analysis of the studies will need to code several specific numerical values. Two common ESs are Cohen's $d\,(M_1 - M_2)/\sigma_{\text{pooled}}$ and the r-index $\sum ZxZy/N$. The r-statistic is usually used with two continuous variables and the d-statistic is used for when one of the variables is dichotomous (Cooper et al., 2000). Rosenthal (1994) describes other potential ES estimates including Z_r, Glass's Δ, Probit d', and Logit d'. Some studies will report an ES; however, others will require the coder to compute the ES. Lipsey and Wilson (2001), and Rosenthal describe different methods for computing *d* and *r* ESs based on available statistics. Most reviewers conducting a meta-analysis of special education will use the *d*-statistic, as did Kroesbergen and Van Luit (2003) in a study of the effects of different mathematics interventions (dichotomous) based on measures of achievement (continuous).

Effect sizes from single-participant designs. Studies involving single-participant designs will need a different statistic. Scruggs and Mastropieri (1998) recommend reviewers use percent of non-overlapping data (PND). PND is the proportion of treatment data points that exceed the highest (or lowest) baseline score. Busk and Serlin (1992) offer an alternative to PND that more closely approximates the *d*-statistic, the standard mean difference (SMD); the SMD is difference between the mean of the baseline points and the mean of the treatment points divided by a measure of variance. A third estimate called the split-middle technique (Kazdin, 1982) looks at the number of treatment data points that fall above (or below) the trend line of the baseline phase. Browder and Xin (1998) used PND to analyze studies of teaching functional reading to students with moderate to severe disabilities. Marquis et al. (2003) used SMD to review studies of positive behavior supports.

Because group designs and single-participant designs use different statistics, they must be analyzed separately. Xin and Jitendra (1999) conducted a meta-analysis of instructional methods for teaching students with LD to solve word problems. The studies they found split almost evenly between group and single-participant designs. Xin and Jitendra coded both types of studies, but analyzed them separately.

Other considerations. Prior to combining ESs to find an overall (mean) ES, many researchers weight the ES. The most frequent weight is the inverse of the variance. Lipsey and Wilson (2001), and Shadish and Haddock (1994) provide formulas for weighting ESs.

If the reviewers do not have enough information to derive an ES statistic, they may still analyze the studies using a vote counting procedure (Bushman, 1994; Cooper et al., 2000). Vote counting procedures can be as simple as comparing the number of studies with positive effects to studies with negative effects. Vote counting may also be used to compute approximate estimates of the *d*-statistic. Cooper et al. provide an example from their own research on the effects of summer school that demonstrates how vote counting can estimate ESs.

Statistical tests of effect sizes. Once the reviewers have ESs, they can begin to analyze them using various statistical tests. The test for homogeneity based on the *Q*-statistic is one analysis (Lipsey & Wilson, 2001). Elbaum, Vaughn, Hughes, Moody, and Schumm (2000), in a study of grouping formats, used the *Q*-statistic to test for differences. Using the homogeneity test, Elbaum et al., found differences between studies of peer tutoring, cross-age tutoring, and cooperative partners. Analysis of variance based on identified factors and regression for ESs are other useful statistical tests (Hedges, 1994). Robinson et al. (1999) used regression to show that coded variables did not have a significant impact on the ES.

Using Synthesis to Develop a Conceptual Framework

As noted previously, the results of integrative reviews are dependent upon the questions asked. Miller and Pollack (1994) state synthesis research can shape theory in three ways: It can (a) give added weight to a theory, (b) highlight variables that limit or change a theory, and (c) test new theories. Miller and Pollock go on to suggest that testing of new theories is a useful, yet rare use of synthesis research. Integrative reviews can highlight gaps in research. Our review demonstrated the dearth of research on all but the five most popular accommodations.

When the effects of many studies show consistent results, the review will provide stronger support for the theory of causality. Kroesbergen and Van Luit (2003) examined research using different instructional methods for teaching mathematics to students with disabilities. They were able to make comparisons of different methods and to demonstrate consistency of effects for direct instruction.

Synthesis research can identify factors that appear to affect variability. McLeskey, Tyler, and Flippin (2004) used synthesis research to look at variability in shortages of special education teachers by location, job description, and diversity of personnel (e.g., ethnicity and languages spoken).

Synthesis research can confirm theory and end lines of research. Conversely, some synthesis research will identify discrepancies in current research and open new lines of future research (Eagley & Wood, 1994).

CONCLUSION

Special education, in its search for effective practices, must rely on synthesis research. Current integrative research has examined the efficacy and effects of special education placements (e.g., Carlberg & Kavale, 1980; Elbaum, 2002), various types of instruction (e.g., Kroesbergen & Van Luit, 2003), and behavioral interventions (e.g., Forness & Kavale, 1996; Robinson et al., 1999). With the added emphasis on empirically founded interventions in schools coming from NCLB and the 2004 amendments to IDEA, we expect to see more frequent uses of synthesis research.

To read integrative reviews, special education practitioners need a basic understanding of the integrative reviews, how they are constructed, what they can tell us, and how to judge their quality.

As the use of integrative reviews increases, researchers will need to continue to produce quality primary research that can be integrated into these reviews. It will also become increasingly important that researchers consider standardizing research protocols. Comparisons of similar interventions are easier when the controls are similar, when researchers use similar experimental designs, and when they report similar statistics.

REFERENCES

Al-Otaiba, S., & Fuchs, D. (2002). Characteristics of children who are unresponsive to early literacy intervention: A review of the literature. *Remedial & Special Education*, *23*, 300–316.

Brigham, F. J., Gustashaw, W. E., Wiley, A. L., & Brigham, M. (2004). Research in the wake of NCLB: Why the controversies will continue and some suggestions for controversial research. *Behavioral Disorders*, *29*, 300–310.

Browder, D. M., & Xin, Y. P. (1998). A meta-analysis and review of sight word research and its implications for teaching functional reading to individuals with moderate and severe disabilities. *Journal of Special Education*, *32*, 130–153.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 193–213). New York: Russell Sage Foundation.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single case research. In: T. R. Kratochwill & J. R. Levin (Eds), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Earlbaum.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research*, *64*, 363–423.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Carlberg, C., & Kavale, K. A. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. *Journal of Special Education*, *14*, 296–309.

Carnine, D. (1997). Bridging the research to practice gap. *Exceptional Children*, *63*, 513–521.

Casto, G., & Mastropieri, M. A. (1986). The efficacy of early intervention programs: A meta-analysis. *Exceptional Children*, *52*, 417–424.

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, *2*, 31–49.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cooper, H. M. (1988). Organizing knowledge synthesis: A taxonomy of literature reviews. *Knowledge in Society*, *1*, 104–126.

Cooper, H. M., & Hedges, L. V. (1994). *The Handbook of research synthesis*. New York: Russell Sage Foundation.

Cooper, H., Valentine, J. C., & Charlton, K. (2000). The methodology of meta-analysis. In: R. Gersten, E. P. Schiller & S. Vaughn (Eds), *Contemporary special education research: Synthesis of the knowledge base on critical instructional issues* (pp. 263–280). Mahwah, NJ: Erlbaum.

Davey, A., Shanahan, M. J., & Schafer, J. L. (2001). Correcting for selective nonresponse in the National Longitudinal Survey of Youth using multiple imputation. *The Journal of Human Resources*, *36*, 500–519.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*, 627–668.

Dorsey, B. L., Nelson, R. O., & Hayes, S. C. (1986). The effects of code complexity and of behavioral frequency on observer accuracy and interobserver agreement. *Behavioral Assessment*, *8*, 349–363.

Dunn, R., Griggs, S. A., Olson, J., Beasley, M., & Gorman, B. S. (1995). A meta-analytic validation of the Dunn and Dunn model of learning-style preferences. *Journal of Educational Research*, *88*, 353–362.

DuPaul, G. J., & Eckert, T. L. (1997). The effects of school-based interventions for attention deficit hyperactivity disorder: A meta-analysis. *School Psychology Review*, *26*, 5–27.

Dush, D. M., Hirt, M. L., & Schroeder, H. E. (1989). Self-statement modification in the treatment of child behavior disorders. *Psychological Bulletin*, *106*, 97–106.

Eagley, A. H., & Wood, W. (1994). Chapter 30. In: H. Cooper, & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 485–500). New York: Russell Sage Foundation.

Elbaum, E. (2002). The self-concept of students with learning disabilities: A meta-analysis of comparisons across different placements. *Learning Disabilities Research & Practice*, *17*, 216–226.

Elbaum, E., Vaughn, S., Hughes, M. T., Moody, S. W., & Schumm, J. S. (2000). How reading outcomes of students with disabilities are related to instructional formats: A meta-analytic review. In: R. Gersten, E. P. Schiller & S. Vaughn (Eds), *Contemporary special education research: Synthesis of the knowledge base on critical instructional issues* (pp. 105–135). Mahwah, NJ: Erlbaum.

Forness, S. R. (2001). Special education and related services: What have we learned from meta-analysis? *Exceptionality*, *9*, 185–197.

Forness, S. R., & Kavale, K. A. (1996). Treating social skill deficits in children with learning disabilities: A meta-analysis of the research. *Learning Disability Quarterly*, *19*, 2–13.

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Lipsey, M. W. (2000). Reading differences between low-achieving students with and with-out learning disabilities: A meta-analysis. In: R. Gersten, E. P. Schiller & S. Vaughn (Eds), *Contemporary special education research: Synthesis of the knowledge base on critical instructional issues* (pp. 81–104). Mahwah, NJ: Erlbaum.

Fuchs, D., Fuchs, L. S., McMaster, K. N., & Al-Otaiba, S. (2003). Identifying children at risk for reading failure: Curriculum-based measurement and the dual-discrepancy approach. In: H. L. Swanson, K. R. Harris & S. Graham (Eds), *Handbook of learning disabilities* (pp. 431–449). New York: Guilford Press.

Fuchs, L. S. (2000). The validity of test accommodations for students with learning disabilities: Differential item performance on reading tests as a function of test accommodations and disability status. Retrieved June 21, 2004, from the Delaware Department of Education Web site: http://www.doe.state.de.us/aab/

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*, 199–208.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, *67*, 67–81.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, *5*, 351–379.

Graham, S., & Harris, K. R. (2003). Students with learning disabilities and the process of writing: A meta-analysis of SRSD studies. In: H. L. Swanson, K. R. Harris & S. Graham (Eds), *Handbook of learning disabilities* (pp. 323–334). New York: Guilford Press.

Hall, J. A., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 17–28). New York: Russell Sage Foundation.

Hartman, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, *10*, 103–116.

Hartman, D. P. (1982). Assessing the dependability of observational data. In: D. P. Hartman (Ed.), *Using observers to study behavior*. San Francisco: Jossey-Bass.

Hedges, L. V. (1994). Fixed effect models. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

Jablonski, B. R., Edgemon, E. A., Wiley, A. L., & Lloyd, J. W. (2005). Large-scale testing accommodations for students with disabilities. Manuscript submitted for publication.

Jensen, P. (2003). What is evidence for evidence-based practice? *Emotional and Behavioral Disorders in Youth*, *3*, 37–48.

Kavale, K. A. (2001a). Decision-making in special education: The function of meta-analysis. *Exceptionality*, *9*, 245–268.

Kavale, K. A. (2001b). Meta-analysis: A primer. *Exceptionality*, *9*, 177–183.

Kavale, K. A., & Forness, S. R. (1983). Hyperactivity and diet treatment: A meta-analysis of the Feingold hypothesis. *Journal of Learning Disabilities*, *16*, 324–330.

Kavale, K. A., & Forness, S. R. (1987). Substance over style: Assessing the efficacy of modality testing and teaching. *Exceptional Children*, *54*, 228–239.

Kavale, K. A., & Forness, S. R. (1996). Social skill deficits and learning disabilities: A meta-analysis. *Journal of Learning Disabilities*, *29*, 226–237.

Kavale, K. A., & Forness, S. R. (2000). Policy decisions in special education: The role of meta-analysis. In: R. Gersten, E. P. Schiller & S. Vaughn (Eds), *Contemporary special education research: Synthesis of the knowledge base on critical instructional issues* (pp. 137–178). Mahwah, NJ: Erlbaum.

Kavale, K. A., & Mattson, P. D. (1983). One jumped off the balance beam: Meta-analysis of perceptual-motor training. *Journal of Learning Disabilities*, *16*, 165–173.

Kazdin, A. E. (1982). *Single-case research designs*. New York: Oxford University Press.

Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial and Special Education*, *24*, 97–114.

Landrum, T. J. (1997). Why data don't matter (guest editorial). *Journal of Behavioral Education*, *7*, 123–129.

Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 111–123). New York: Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lloyd, J. W., Forness, S. R., & Kavale, K. A. (1998). Some methods work better than others. *Intervention in School and Clinic*, *33*, 195–200.

Lloyd, J. W., Pullen, P. C., Tankersley, M., & Lloyd, P. A. (2006). Critical dimensions of experimental studies and research syntheses that help define effective practices. In: B. G. Cook & B. R. Schirmer (Eds), *What is special about special education*. Austin, TX: ProEd.

Marquis, J. G., Horner, R. H., Carr, E. G., Turnball, A. P., Thompson, M., Behrens, G. A., Magito-McLaughlin, D., McAtee, M. L., Smith, C. E., Ryan, K. A., & Doolabh, A. (2003). A meta-analysis of positive behavior support. In: R. Gersten, E. P. Schiller & S. Vaughn (Eds), *Contemporary special education research: Synthesis of the knowledge base on critical instructional issues* (pp. 137–178). Mahwah, NJ: Erlbaum.

Mastropieri, M. A., & Scruggs, T. E. (1989). Constructing more meaningful relations: Mnemonic instruction for special populations. *Educational Psychology Review*, *1*, 83–111.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research syntheses. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

McLeskey, J., Tyler, N. C., & Flippin, S. S. (2004). The supply of demand for special education teachers: A review of research regarding the chronic shortage of special education teachers. *The Journal of Special Education*, *38*, 5–21.

Mertens, D. M. (2005). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (2nd ed.). Thousand Oaks, CA: Sage.

Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 457–483). New York: Russell Sage Foundation.

Mostert, M. P. (2001). Characteristics of meta-analyses reported in mental retardation, learning disabilities, and emotional and behavioral disorders. *Exceptionality*, *9*, 199–225.

Mostert, M. P., & Kavale, K. A. (2001). Evaluation of research for usable knowledge in behavioral disorders: Ignoring the irrelevant, considering the germane. *Behavioral Disorders*, *27*, 53–68.

Murawski, W. W., & Swanson, H. L. (2001). A meta-analysis of co-teaching research: Where are the data? *Remedial and Special Education*, *22*, 258–287.

National Institute of Child Health and Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidenced based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: U.S. Government Printing Office (NIH Publication No. 00–4769).

Nelson, J. R., Roberts, M. L., & Mathur, S. R. (1999). Has policy exceeded knowledge base? A review of the functional behavioral assessment literature. *Behavioral Disorders*, *24*, 169–179.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*, 137–148.

Orwin, R. G. (1994). Evaluating coding decisions. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 139–162). New York: Russell Sage Foundation.

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 163–175). New York: Russell Sage Foundation.

Quinn, M. M., Kavale, K. A., Mathur, S. R., Rutherford, R. B., & Forness, S. R. (1999). A meta-analysis of social skill interventions for students with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders*, *7*, 54–64.

Reid, R., Gonzalez, J. E., Nordness, P. D., Trout, A., & Epstein, M. H. (2005). A meta-analysis of the academic status of students with emotional/behavioral disturbance. *Journal of Special Education*, *38*, 130–143.

Robinson, T. R., Smith, S. W., Miller, M. D., & Brownell, M. T. (1999). Cognitive behavior modification of hyperactivity-impulsivity and aggression: A meta-analysis of school-based studies. *Journal of Educational Psychology*, *91*, 195–203.

Rosenthal, R. (1994). Parametric measures of effect size. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Sabornie, E. J., Culuinan, D., Osborne, S., & Brock, L. B. (2005). Intellectual, academic, and behavioral functioning of students with high-incidence disabilities: A cross-categorical meta-analysis. *Exceptional Children*, *72*, 47–63.

Sasso, G. M., Conroy, M. A., Stichter, J. P., & Fox, J. J. (2001). Slowing down the bandwagon: The misapplication of functional behavior assessment for students with emotional or behavioral disorders. *Behavioral Disorders*, *26*, 282–296.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification*, *22*, 221–242.

Seethaler, P. M., & Fuchs, L. S. (2005). A drop in the bucket: Randomized controlled trials testing reading and math interventions. *Learning Disabilities Research & Practice*, *20*, 98–102.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect sizes. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Sherman, R. P. (2000). Tests of certain types of ignorable nonresponse in surveys subject to item nonresponse or attrition. *American Journal of Political Science*, *44*, 362–374.

Stage, S. A., & Quiroz, D. R. (1997). A meta-analysis of interventions to decrease disruptive classroom behavior in public education settings. *School Psychology Review*, *26*, 333–368.

Stock, W. A. (1994). Systematic coding for research synthesis. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 125–138). New York: Russell Sage Foundation.

Swanson, H. L. (1999). Reading research for students with LD: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities*, *32*, 504–532.

Swanson, H. L., & Hoskyn, M. (2001). Instructing adolescents with learning disabilities: A component and composite analysis. *Learning Disabilities Research & Practice*, *16*, 109–119.

Swanson, H. L., & Malone, S. (1992). Social skills and learning disabilities: A meta-analysis of the literature. *School Psychology Review*, *21*, 427–443.

Talbott, E., Lloyd, J. W., & Tankersley, M. (1994). Effects of reading comprehension interventions for students with learning disabilities. *Learning Disability Quarterly*, *17*, 223–232.

Vaughn, S., & Dammann, J. E. (2001). Science and sanity in special education. *Behavioral Disorders*, *27*, 21–29.

Weiss, M. P., & Brigham, F. J. (2000). Co-teaching and the model of shared responsibility: What does the research support? In: T. E. Scruggs & M. A. Mastropieri (Eds), *Educational interventions: Advances in learning and behavioral disabilities* (Vol. 14, pp. 217–245). Stamford, CT: JAL.

White, W. A. T. (1988). A meta-analysis of effects of direct instruction in special education. *Education and Treatment of Children*, *11*, 364–374.

Wortman, P. M. (1994). Judging research quality. In: H. Cooper & L. V. Hedges (Eds), *The handbook of research synthesis* (pp. 97–109). New York: Russell Sage Foundation.

Xin, Y. P., & Jitendra, A. K. (1999). The effects of instruction in solving mathematical word problems for students with learning problems: A meta-analysis. *Journal of Special Education*, *32*, 207.

MATH DISABILITIES: A PRELIMINARY META-ANALYSIS OF THE PUBLISHED LITERATURE ON COGNITIVE PROCESSES

Lee Swanson and Olga Jerman

ABSTRACT

This chapter synthesized some of the published literature comparing the cognitive functioning of children with math disabilities (MD) with (1) average achieving children, (2) children with reading disabilities (RD), and (3) children with comorbid disabilities (RD+MD). Twenty-one studies, which yielded 194 effect sizes (ESs), indicated that average achievers outperformed children with MD on measures of verbal problem solving ($M = -0.58$), naming speed ($M = -0.70$), verbal ($M = -0.70$) and visual-spatial working memory (WM, $M = -0.63$), and long-term memory (LTM, $M = -0.72$). The results further indicated that children with MD outperformed children with combined disabilities on measures of literacy ($M = 0.75$), visual-spatial problem solving ($M = 0.51$), LTM ($M = 0.44$), short-term memory (STM) for words ($M = 0.71$), and verbal WM ($M = 0.30$). Children with MD could only be clearly differentiated from children with RD on measures of naming speed (-0.23) and visual-spatial WM (-0.30). The magnitude of ESs was persistent across age and severity of math disability. Hierarchical linear modeling (HLM)

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 285–314

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19011-4

indicated that the magnitude of ES in overall cognitive functioning between MD and average achievers was due to verbal WM deficits when the effect of all other variables (e.g., age, IQ, reading level, other domain categories) were partialed out. The results are discussed within the context of defining MD by level of severity of WM abilities.

MATH DISABILITIES: A PRELIMINARY META-ANALYSIS OF THE PUBLISHED LITERATURE ON COGNITIVE PROCESSES

The purpose of this chapter is to provide a quantitative synthesis of the published literature comparing children with math disabilities (MD) to either average achievers or those children who are suffering comorbid disabilities (i.e., reading disabilities (RD)) on various cognitive measures. Although not a quantitative analysis, one of the most comprehensive syntheses of the cognitive literature on MD was provided by Geary (1993; also see Geary, 2004, for a review). His review indicated that children with MD are a heterogeneous group and show one of the three types of cognitive disorders.

One type disorder characterizes children with MD as suffering visual-spatial difficulties. These children have difficulties representing numerical information spatially. Example difficulties represent misalignment of numerals in multi-column arithmetic problems and rotation of numbers. Further, they have difficulties in areas that require spatial ability such as geometry and place values. Recent work by Geary, Hoard, Byrd-Craven, and DeSoto (2004) suggests that these deficits are not due to poor spatial abilities, but rather to poor monitoring of the sequence of steps of an algorithm and from poor skills in detecting and then self-correcting errors.

Another type of math disorder is procedural. Children in this category generally use developmentally immature procedures in numerical calculations and therefore have difficulties in sequencing multiple steps in complex procedures. For example, Gross-Tsur, Manor, and Sha1ev (1996) indicated that children with MD have a basic understanding of number and small quantities. However, children with MD have difficulties keeping information in working memory (WM) and monitoring the counting process (Hitch & McAuley, 1991) that creates errors in their counting. Other studies (e.g., Jordan, Hanich, & Kaplan, 2003a, b; Jordan & Montani, 1997) indicate that children with MD have difficulties in solving simple and complex arithmetic problems. These differences are assumed to involve both procedural and memory-based deficits. Procedural deficits relate to miscounting or losing track of the counting process.

A final disorder characterizes children with MD as suffering from a semantic memory deficit. These children are characterized as having weak fact retrieval and high-error rates in recall. Disruptions in ability to retrieve basic facts from long-term memory (LTM), due to inhibition, may be a defining feature of MD (Geary, 1993). Further, Geary's review suggested that the characteristics of these retrieval deficits, such as slow solution times, suggest that children with MD do not suffer from simple developmental delay, but rather from a more persistent cognitive disorder across a broad age span.

Regardless of the type of disorder, the majority of these studies suggest that children with MD suffer memory deficits. For example, the literature suggests that children with MD do not show the shift from direct counting procedures to a memory-based production of the solution (e.g., Swanson & Rhine, 1985). That is, they do not remember that certain combinations of new numbers yield a certain result and have difficulty accessing facts from LTM and therefore have difficulty engaging in labor-intensive calculations. Geary (1993, 1994) has also suggested that memory representations for arithmetic facts are supported in part by the same phonological and semantic memory systems that support decoding and reading comprehension. If this is the case, then perhaps phonological processes that contribute to reading disorders might also be a source of math retrieval difficulties in children with mathematical disorders. This co-occurrence between math and reading has been assumed (e.g., Hecht, Torgesen, Wagner, & Rashotte, 2001).

There are several reasons why it is difficult to determine from existing literature whether cognitive processing of children with MD is distinct from other children, especially those with RD. First, the operational criteria for measures used in the selection of children with MD and RD vary across studies. For instance, measures used to establish MD vary from the 48th percentile to the 8th percentile. Geary further indicated that like reading disorders there really is no universally agreed upon criteria for the diagnosis of math disorders. Variations in definitions and issues of comorbidity have raised questions about whether some of the processes associated with MD include cognitive sub-processes specific to math or whether deficits affecting math extend to other domains, such as reading. More specifically, there have been a range of terms used to define MD and these have varied on different criteria. Geary included children who fall below the 30th percentile (Geary, Hoard, & Hamson, 1999) or the 35th percentile (Geary, Hamson, & Hoard, 2000). Jordan and colleagues referred to children with math difficulties as children below the 35th percentile. Koontz and Berch (1996) use scores falling below the 25th percentile as their criteria for MD performance on a group administered test.

Regardless of the cut-off score for determination of MD, the most generally agreed upon failure of children with MD is in learning to remember arithmetic facts (see Geary, 1993; however, see Landerl, Bevan, & Butterworth, 2004). Poor recall of arithmetic facts, of course lead to difficulties executing calculating procedures and immature problem solving strategies (Geary, 1993). It is important to note in many of these comparison studies, that the classification procedures are not that distinct (orthogonal) from the comparison measure. For example, MD and nonMD children are compared on computation and word problem solving measures when the classification measures of MD themselves (e.g., standardized tests) include similar mathematical operations. Thus, it is not surprising that many of the children with MD are characterized as having primary deficits in calculation.

In this meta-analysis, we attempt to look at those cognitive processes that are independent of the classification measures. We attempt to answer three questions about the literature.

1. Are cognitive deficits in children with MD distinct from their average achieving counterparts and children with comorbid disorders (e.g., RD)?
2. Are the cognitive deficits a function of variations in age? The majority of studies that have compared children with MD have focused on the elementary grades. However, we would like to determine if some of the same deficits emerge in studies that include older participants.
3. Do the cognitive deficits that emerge in children with MD vary as a function of definitional criteria? We compare studies on cognitive outcomes as a function of severity of the MD and intelligence level.

METHOD

Identification of Studies

Several approaches were used to locate the relevant studies. The principle method of location involved a computer search of the PsycINFO database. The search used the following terms: math disabled, math disabilities, dyscalculia, less skilled math, math disabled/reading disabled, arithmetic disabled, poor problem solvers, problem solving in math, and problem solving and math. From these articles, lists of primary researchers were developed. Thus, we further accessed articles published by the following authors' last names: Geary, Jordan, Fletcher, Fuchs, and Siegel. Second, a manual search was conducted of journals where the majority of articles were published (e.g., *Journal of Learning Disabilities*, *Journal of Experimental Child Psychology*). Finally, a hand search was done on all studies cited in the aforementioned articles. In sum, the sample search comprised articles published between 1970 and June 2003. Collectively, these methods identified over 300 articles. The pool of literature was then narrowed down to 85 potentially relevant studies based on selecting only comparative studies (e.g., children with MD compared with a nonMD group). Some of the reasons for rejection of articles is provided in the appendix.

Article Inclusion Criteria

Eighty-five (85) 'potential studies' were further evaluated to determine their relevance to the current review. To be included in the meta-analysis each study had to satisfy the following criteria:

1. An MD group was compared to a nonMD group (non indication of MD). Other comparison groups (e.g., children with RD, children with ADHD) were also coded if an MD group was in the sampling.
2. Within the MD groups, at least one math subgroup had no reported comorbidity (e.g., RD, ADHD).
3. Each study provided a standardized measure of intelligence separated by group.
4. Each study reported scores from a standardized mathematics assessment separated by comparison group.

Several studies were excluded if (a) they were not published in refereed journals (they were book chapters or dissertations); (b) they failed to provide enough quantitative data to calculate the ESs; (c) they failed to include a comparison group; and/or (d) they failed to provide information of ability group performance on a standardized math and/or IQ test. Some of the articles excluded from the meta-analysis and the reasons for exclusion can be provided by the authors.

Overall, 28 studies met the inclusion criteria and were retained for the final inclusion into the present meta-analysis (denoted by asterisk in the references). The psychometric characteristics of the comparisons are provided in Table 2.

Coding Procedure

Each study was coded for the following information: (a) sample characteristics, (b) classification measures, and (c) cognitive measures.

Attributes of the Study

Each study provided (a) the year of the study; (b) the name of the first author; (c) the number of coauthors, and (d) the country where the study was carried out.

Attributes of the Participants

There were four identified subgroups: normal achieving control group; math disabled; reading disabled; math and reading disabled. According to the inclusion criteria each study provided at least one MD and one nonMD comparison group. Other attributes of the participants coded included (b) the number of participants in each subgroup; (c) the number of males in each subgroup; (d) the mean age of the group (converted into months); and (e) participants' primary language. Studies were also coded for the (f) socioeconomic status (SES) and (g) ethnicity status.

Comparison Measures

All classification measures (IQ, mathematics, reading) were converted to standard scores. In those cases when only a range was reported, a middle value was assigned. In terms of comparative measures, cognitive tasks were initially organized into 17 categories: language, comprehension, speed/rapid naming, phonological processing, math-problem solving, vocabulary, reading, visual-spatial tasks, nonverbal-(visual-spatial) problem solving, fine motor/gross motor/visual motor tasks, writing and spelling, LTM, short-term memory (STM) (words), STM (digits), WM (verbal), WM (visual), and attention (behavior ratings). Because of small number of ESs, these 17 categories were further aggregated into 10 broader domains. Although every attempt was made to separate out tasks related to the classification variable (in case of MD-arithmetic calculation, in the case of RD-word recognition), some of the categories were closely related (literacy, problem solving-verbal). Regardless, the following 10 categories were measures not used in the classification of the sample.

1. *Literacy-reading.* The majority of dependent measures in this domain included reading comprehension, writing, vocabulary, and phonological awareness. Tasks presented within this domain required reading, listening and comprehension vocabulary, phonological processing (e.g., phonemic deletion task), spelling, and recognizing visual form of words or sounds.
2. *Problem solving-verbal.* This domain included measures of accuracy in solving story problems.
3. *Speed.* This domain included measures of the rapid naming of letters, numbers and objects, and speed measures such as coding.
4. *Problem solving-visual-motor.* This domain included measures that required the manual manipulation of objects (blocks, discs, puzzles) to solve a problem (e.g., Tower of Hanoi).
5. *Long-Term memory.* This domain included measures that tapped previous knowledge or memory for general information (e.g., answer questions-what is capital of California, recall a story they heard).
6. *STM-words.* This domain included tasks that required the recall of increasingly difficult sets of words and letters. This domain varied from verbal WM (below) in that no distracter question was asked of the participant prior to retrieval.
7. *STM-numbers.* This domain included tasks that required the recall of increasingly difficult sets of digits. This domain varied from WM in that no distracter question was asked of the participant prior to retrieval.
8. *WM-verbal.* This domain included tasks that required the recall of increasingly difficult sets of words and sentences. This domain varied from verbal STM in that process and storage components were included. An example of a verbal WM test was a semantic association task in which a child was presented a set of words reflecting different categories (word sets range from 2 to 9 monosyllabic words). Before recalling the words, however, the participant was asked whether a particular word or word category was included in the set.
9. *WM-visual-spatial.* This domain included tasks that required the recall of increasingly difficult sets of dots, designs, and objects. An illustration of spatial WM was Visual Matrix task in which a participant was presented with a series of dots in a matrix and was allowed 5 s to study the matrix. The matrix was then removed and the participant was asked a process question (e.g., Are there any dots in the first column?). After answering the process question the child was asked to draw the dots in the correct boxes on the blank matrix.
10. *Attention.* This domain included observations that focused mostly on classroom behavior for which measures of attention or behavior were

recorded. For example, in one study Conners' Continuous Performance Test was used to assess sustained attention; in another study a developmental questionnaire provided information on participants' activity level, impulsivity, attention, and inattention.

Calculation of Effect Sizes

For each measure an ES was computed (Cohen's *d*, 1988) and was then weighted by the reciprocal in the sampling variance. The dependent measure for the estimate of effect size (ES) was defined as est $= (d/(1/v))$, where d (Mean of MD−Mean of comparison group/average of standard deviation for both groups), and v is the inverse of the sampling variance, $v = (N_{\text{md}} + N_{\text{nmd}})/(N_{\text{md}} \times N_{\text{nmd}}) + d^2/[2(N_{\text{md}} + N_{\text{nmd}})]$. Means and standard deviations were used in the computation of 98% of the ESs. In the remaining cases, *F*- and *t*-ratios were converted to ESs. Cohen's criterion was used for the interpretation of the magnitude of the ESs. According to Cohen's criterion, an ES of 0.20, in absolute value, is considered small, ESs of 0.50 and 0.80, in their absolute values, are considered moderate and large, respectively. As suggested by Hedges and Olkin (1985), outliers were removed from the analysis of main effects. Outliers were defined as ESs lying beyond the first gap of at least one standard deviation between adjacent ES values in a positive direction (Bollen, 1989). Ten ESs were removed from the analysis.

Statistical Analysis

The analysis of each category of measure reported separately is shown in Table 3. For the category of each dependent measure, a homogeneity statistic Q was computed to determine whether separate ESs within each category shared a common ES. The statistic Q has a distribution similar to the distribution of χ^2 with k-1 degrees of freedom, where k is the number of ESs. A significant χ^2 indicated that the study features significantly moderated the magnitude of ESs. If the homogeneity was not achieved, then the influence of outliers was assessed using a 95% confidence interval.

Hierarchical linear modeling (HLM) was employed to test the hypothesis that age, intelligence, math level, and/or type of cognitive measure influenced the magnitude of the ES (e.g., Bryk & Raudenbush, 1992). One advantage of HLM over traditional methods of analyzing ESs was that multiple measures within studies do not have to be averaged (aggregated within studies) or collapsed. Another advantage is it allows for the assessment of the extent to which individual studies and variation within studies influence outcomes. Further, HLM can accommodate incomplete data and iteratively solve for coefficients at two levels, which are calculated simultaneously. Level 1 equations represented the level of the ES for each observation (K). Level 2 was ES differences between studies that served to predict Level 1 coefficients for the intercept and slope.

In the present study, we first calculated an unconditional model and then two conditional models (models that attempt to identify variables that significantly moderate ESs) using the SAS PROC MIXED program (SAS, 1999). The unconditional model can be viewed as a one way random effects ANOVA model. This model has one fixed effect, intercept, and two variance components. One variance component represented the variation between the studies and the other represented the variations among the ESs within the studies. Random effects were also assessed. Random effects variance was defined as variance from a true ES. Random effects can be viewed as the variance of the true ESs in a population of studies from which the synthesized studies constituted a random sample.

In the first conditional model, the intercept was a dependent variable and was used to predict individual ESs related to the various classification measures (e.g., IQ, math, reading). Specifically, the first conditional model tested whether the dependent variable (ES difference between MD and average achiever) was a function of variations in the classification variables (IQ, math, reading, age) and random error. In the second conditional model, we determined if specific cognitive domains (e.g., WM) moderated overall cognitive functioning.

When one or more predictors are introduced into the HLM model, the reductions in magnitude of the various variance components are analogous to ESs. This is similar to the use of R^2 in linear regression. The primary distinction between linear regression and HLM is that several R^2 values are relevant in HLM because there are several variance components (Snijders & Bosker, 2003). The intraclass correlation coefficient separates the total variability into within study and between study variance.

Interrater Agreement

Three doctoral students coded studies; then a fourth doctoral student evaluated a randomly selected subset of the articles for reliability of coding. The overall structure of the coding system yielded a reliable percentage of interrater agreement across all codes ($>90\%$ agreement).

RESULTS

Study Characteristics

Table 1 provides an overview of the characteristics of each study in the meta-analysis and shows the study's publication year, the journal, the primary author, country where study occurred, sample size, and mean age of the sample. Articles were published most frequently in the *Journal of Experimental Child Psychology*, *Journal of Clinical & Experimental Neuropsychology*, *Journal of Learning Disabilities*, *and Learning Disability Quarterly*. Publication dates ranged from 1983 to 2002 with the average year of publication 1994. The number of authors ranged from 1 to 9. Out of 29 studies, 17 were conducted in the USA, 5 in Canada, 3 in Italy, 2 in Spain, and 1 in New Zealand.

SES status of the participants was reported in 8 studies and the ethnic background was given in 9 studies. Twenty-four studies indicated the ratio of males to females in participant selection. However, no study separated the math performance as a function of gender, ethnicity, or SES. Therefore, math performance as a function of gender, ethnicity, and/or SES was not compared across the studies. Table 2 provides an overview on the psychometric information (IQ, math, and reading) on participants for three comparisons MD vs. nonMD, MD vs. RD, and MD vs. comorbid group (RD + MD). Also provided are the ESs for IQ, math and reading between the groups.

Domain Categories

Table 3 provides the weighted means and standard deviations for ESs for each category and comparison. Prior to the analysis, naming speed measures were corrected for the direction of ES so they could be combined with measures of accuracy and rate. As shown on Table 3, there were 194 dependent measures related to comparisons between MD and average achievers, which yielded a mean ES of −0.52. Using Cohen's criterion, −0.52 was considered an ES in the moderate range. Moderate ESs (0.50–0.80) emerged across several categories such as verbal and visual problem solving, speed, LTM, STM for words, verbal WM, and visual-spatial WM. No large ES in the range of 0.80 or better emerged. We compared whether the ESs as a function of category were significantly different. For the weighted ES, a significant effect was found for domain, χ^2 (9, $N = 193$) = 77.73, $p < 0.001$. A Scheffé test indicated that the negative ESs were significantly ($p < 0.05$)

Table 1. Study Characteristics.

Reference	Journal	Country	Sample Size MD Group	Age MD	Sample Size Control group	Age Control Group
Garnett and Fleischner (1983)	*Learning Disability Quarterly*	USA	120	125.5	120	125.5
Lund, Hall, Wilson, and Humphreys (1983)	*Journal of Experimental Child Psychology*	USA	12	95.28	32	97.44
Nolan, Hammeke, and Barkley (1983)	*Journal of Clinical Child Psychology*	USA	12	129.33	12	119.75
Fletcher (1985)	*Journal of Experimental Child Psychology*	USA	13	120.46	16	122
Share, Moffit, and Silva (1988)	*Journal of Learning Disabilities*	New Zealand	30	156	390	156
Siegel and Ryan (1988)	*Developmental Psychology*	Canada	73	126	138	126
Siegel and Ryan (1989)	*Child Development*	Canada	36	134.15	74	116.3
Loveland, Fletcher, and Bailey (1990)	*Journal of Clinical & Experimental Neuropsychology*	USA	12	137.1	14	131.8
Mattson, Sheer, and Fletcher (1992)	*Journal of Clinical & Experimental Neuropsychology*	USA	8	144	10	148.8
Montague and Applegate (1993)	*Journal of Special Education*	USA	30	164.4	30	154.8
Lennox and Siegel (1993)	*Applied Psycholinguistics*	Canada	140	127.5	81	127.5
Swanson (1993)	*Journal of Experimental Child Psychology*	USA	19	120.12	38	122.8
Brookshire, Butler, Ewing-Cobbs, and Fletcher (1994)	*Journal of Clinical & Experimental Neuropsychology*	USA	10	4–7 grade	20	4–7 grade

Table 1. (*Continued*)

Reference	Journal	Country	Sample Size MD Group	Age MD	Sample Size Control group	Age Control Group
Miles and Stelmack (1994)	*Journal of Clinical & Experimental Neuropsychology*	Canada	8	129.6	10	140.4
Shafrir and Siegel (1994)	*Journal of Learning Disabilities*	Canada	88	304.8	130	352.8
Swanson (1994)	*Learning Disabilities Research & Practice*	USA	26	129.36	47	129.36
Lucangeli, Coi, and Bosco (1997)	*Learning Disabilities Research & Practice*	Italy	30	5 grade	30	5 grade
Badian (1999)	*Annals of Dyslexia*	USA	25	56.4	107	57.6
Geary et al. (1999)	*Journal of Experimental Child Psychology*	USA	15	83	35	81
Gonzalez and Espinel (1999)	*Learning Disability Quarterly*	Spain	44	93.72	60	93.72
Passolunghi, Cornoldi, and De Liberto (1999)	*Memory & Cognition*	Italy	15	115.2	18	115.2
Geary et al. (2000)	*Journal of Experimental Child Psychology*	USA	12	83	26	81
Lindsay, Tomazic, Levine, and Accardo (2001)	*Journal of Developmental & Behavioral Pediatrics*	USA	42	150.6	107	150.6
Mazzocco (2001)	*Journal of Learning Disabilities*	USA	34	73.6	165	72.4
Passolunghi and Siegel (2001)	*Journal of Experimental Child Psychology*	Italy	23	112.8	26	112.8
Gonzalez and Espinel (2002)	*Learning Disability Quarterly*	Spain	60	93.72	44	93.72
Klorman et al. (2002)	*Biological Psychiatry*	USA	9	104.4	28	123.6
Sikora, Haley, Edwards, and Butler (2002)	*Developmental Neuropsychology*	USA	17	142.8	16	144

Table 2. Psychological and Demographic Information on Participants.

	Chronological Age Matched ($N = 784$)			Math Disabled ($N = 527$)			Effect Size	
	M	SD	Range	M	SD	Range	M	SD
Age	124.51	52.22	72–158	123.64	58.93	73–157	0.12	0.34
IQ	105.59	8.70	80–120	99.69	8.51	80–119	–0.59	0.40
Math	105.64	6.51	96–119	84.76	5.93	75–96	–2.19	1.13
Reading	106.80	5.93	96–113	98.37	7.68	87–109	–0.59	0.49
	Reading disabled ($N = 224$)			Math disabled ($N = 250$)			Effect size	
Age	131.25	82.30	59–141	135.34	76.84	59–142	0.08	0.33
IQ	96.78	7.55	80–107	97.83	8.93	80–105	–0.31	0.56
Math	95.75	8.87	85–103	86.61	6.56	75–87	–1.11	1.50
Reading	80.69	6.75	66–87	99.058	8.16	97–1054	2.27	1.16
	Comorbid (MD & RD) ($N = 135$)			Math disabled ($N = 294$)			Effect size	
Age	122.49	47.49	57–322	135.76	83.65	56–304	–0.54	0.99
IQ	92.43	5.10	89–98	99.92	5.64	94–112	0.59	0.16
Math	84.65	2.84	81–87	86.02	7.92	75–88	0.26	0.24
Reading	82.83	1.25	81–87	100.75	8.28	84–108	1.68	0.47

Note: Negative effect size is in favor of contrast group and positive effect size is in favor of MD group.

larger for LTM, speed, and verbal WM when compared to the other categories (LTM = naming speed = verbal WM = visual-spatial WM > problem solving-verbal = problem solving-visual = STM-words > STM-digits = literacy > attention).

As also shown in Table 3, there were approximately 58 ESs in which we could establish comparisons between MD and RD children. These dependent measures were averaged and yielded a mean ES of –0.10. As shown across all categories, the magnitudes of ESs were low between the two groups. Positive ESs indicated that the MD actually did slightly better than the RD on some literacy measures (e.g., literacy, problem solving-verbal, STM-word). This finding made sense to us since mathematical reasoning and problem solving are strongly related to measures of literacy (e.g., reading comprehension, Swanson, Cooney, & Brock, 1993). The advantages, ESs in the low range, were found for RD children when compared with MD children on measures of speed (naming speed) and visual-spatial WM. When comparing the weighted ESs, a significant effect was found for domain, χ^2

Table 3. Weighted Effect Sizes, Standard Error, Confidence Intervals, and Homogeneity of Categories for Comparisons between MD and nonMath Disabled (MD/NMD), MD and Reading Disabled (MD/RD), and MD and RD + MD (CMOR) (Corrected for Outliers).

Comparison	*K*	Effect Size	Standard Error	Lower	Upper	Homogeneity *Q*
Total across categories						
MD/NMD	194	−0.52	0.01	−0.56	−0.48	767.05***
MD/RD	58	−0.10[a]	0.03	−0.16	−0.04	263.35***
MD/CMOR	102	0.26[a]	0.02	0.22	0.31	650.86***
1. Literacy (vocabulary, reading comprehension)						
MD/NMD	19	−0.30	0.05	−0.40	−0.40	73.52***
MD/RD	6	0.11	0.07	−0.02	0.25	2.00
MD/CMOR	10	0.75	0.06	0.62	0.88	49.30**
2. Problem solving-verbal						
MD/NMD	29	−0.58	0.04	−0.67	−0.49	242.41***
MD/RD	1	0.10	–	–	–	–
MD/CMOR	15	0.13	0.05	0.02	0.23	107.72
3. Speed-naming						
MD/NMD	17	−0.70	0.06	−0.83	−0.56	55.70***
MD/RD	6	−0.23	0.13	−0.49	0.02	0.38
MD/CMOR	10	−0.39	0.09	−0.58	−0.19	6.01
4. Visual-spatial problem solving						
MD/NMD	23	−0.48	0.05	−0.47	−0.31	41.61***
MD/RD	4	0.04	0.09	−0.17	0.18	8.90*
MD/CMOR	10	0.51	0.06	0.38	0.64	44.03***
5. LTM-retrieval (e.g., general information)						
MD/NMD	15	−0.72	0.09	−0.90	−0.54	35.49***
MD/RD	–	–	–	–	–	–
MD/CMOR	9	0.44	0.12	0.20	0.69	10.35
6. STM-words						
MD/NMD	16	−0.45	0.06	−0.58	−0.32	44.78***
MD/RD	30	0.16	0.13	−0.10	0.42	7.33*
MD/CMOR	4	0.71	0.12	0.46	0.96	12.61**
7. STM-digits/numbers						
MD/NMD	11	−0.26	0.07	−0.41	0.10	48.94***
MD/RD	4	0.03	0.14	−0.24	0.32	6.35
MD/CMOR	9	−0.08	0.11	−0.30	0.13	110.57***
8. WM-verbal						
MD/NMD	43	−0.70	0.04	−0.79	−0.61	83.84***
MD/RD	19	−0.07	0.06	−0.19	0.04	139.95***
MD/CMOR	20	0.30	0.06	0.17	0.42	86.49**

Table 3. (*Continued*)

Comparison	K	Effect Size	Standard Error	Lower	Upper	Homogeneity Q
9. WM-visual spatial						
MD/NMD	13	−0.63	0.07	−0.77	−0.48	28.14^{**}
MD/RD	13	−0.30	0.07	−0.44	−0.16	35.43^{**}
MD/CMOR	13	0.23	0.07	0.08	0.38	14.10
10. Attention						
MD/NMD	8	−0.15	0.09	−0.33	0.03	34.83^{***}
MD/RD	0	–	–	–	–	–
MD/CMOR	2	−0.57	0.11	−0.79	−0.35	6.97^{*}

Note: MD = Math Disabled only, NMD = nonmath disabled-average achiever, RD = reading disabled, CMOR = comorbid group with both low reading and math; K = number of measures, Lower and Upper = 95% level of confidence range.
[a]Positive effect sizes favor MD and negative effect sizes favor comparison group;
$^{*}p<05$;
$^{**}p<0.01$;
$^{***}p<0.001$.

(8, $N = 57$) = 57.17, $p<0.001$. A Scheffé test indicated that ESs were significantly higher (positive) for literacy and problem solving and negative for visual-spatial WM and speed (literacy = visual-spatial WM = problem solving-verbal = problem solving-visual = STM-words = STM-digits = verbal WM > speed = visual-spatial WM). However, these overall results should be interpreted with caution because of the infrequent number of ESs.

As shown in Table 3, a comparison was also made between MD and children who had both reading and math problems. One hundred and two dependent measures were averaged and yielded a mean ES of 0.26. Positive ESs in Table 3 indicated that the children with MD did better than the comorbid group. As shown in Table 3, MD children did better than the comorbid group (moderate ES range) on measures of literacy, visual-problem solving, LTM, and STM for words. MD children were inferior to the comorbid group on measures of attention and speed. For the weighted ES, a significant ($p<0.05$) effect was found for domain, χ^2 (9, $N = 81 =$ 234.86, $p<0.001$. A Scheffé test indicated that ESs were significantly larger ($p<0.05$) for the literacy, problem solving-visual and STM for words (literacy = visual-spatial problem solving = STM-words > verbal WM = visual-spatial WM > speed > attention).

We further explored whether the cognitive variables were correlated with the classification measures (standard scores in IQ, reading, and Math).

Table 4. Correlations of Categorical Variables and Age with Total Effect Size (MD/NMD) across Domains, Effect Size for IQ, Reading, and Math ($N = 194$).

	Total ES	ES – IQ	ES – Math	ES – Reading
Age	0.04	−0.11	−0.35***	−0.001
Verbal WM	−0.13	0.19*	−0.24***	0.26***
Visual-spatial WM	−0.03	−0.20*	−0.12	−0.01
STM-words	−0.02	0.36***	0.09	0.16
STM-digits	0.09	−0.21**	−0.22**	−0.10
LTM	−0.04	0.04	0.25**	0.14
Speed	−0.09	−0.20*	0.02	−0.08
Problem solving	−0.03	−0.12	0.19*	−0.11

Note: WM = all working memory measures, STM = all short-term memory measures, Problem solving = all verbal and visual-spatial problem solving measures.
*$p<0.05$,
**$p<0.01$,
***$p<0.001$.

Because the number of ESs was more frequent when comparing MD from average achievers that other group comparisons, these ESs were studied further. Table 4 shows the correlation between the ESs of children with MD and average achievers as a function of the domain categories. Because of the infrequent number of ESs, certain domains were collapsed into more general categories. These variables were then dummy coded (1 represented the category and 0 was the comparison to all other categories that did not include this particular domain). As shown in Table 4, ESs for domains were correlated with the MD samples age, total ESs across the various categories, and ESs comparing MD and normal achieving nonMD on the classification measures (IQ, math, and reading). Significant coefficients emerged between specific domains (WM, STM) and ESs for intelligence, between specific domains (verbal WM, STM-digits, LTM, problem solving) and ESs for math, and verbal WM and ESs for reading. The mean chronological age of the MD group correlated significantly with math ESs.

Table 5 shows a comparison of MD/NMD ESs as a function of age and the degree of severity in MD. Two subgroups of MD were formed: severe MD (studies reporting mean standard scores on the MD sample below 89) and moderate MD (studies reporting standard scores >89). For comparisons as a function of age, studies were also separated into those studies that included a mean sample of children at or above 8.5 years old from those below 8.5 years of age. When comparing MD/NMD weighted mean ESs, a

Table 5. Effect Size as Function of Severity of Math Disability and Age.

	Severe MD		Moderate MD	
	M	SD	M	SD
Younger	−0.44 ($K = 49$)	2.49	−0.54 ($K = 20$)	1.39
Older	−0.62 ($K = 64$)	2.03	−0.53 ($K = 61$)	1.63

Note: $K =$ number of dependent measures.

significant effect was found for age χ^2 (1, $N = 193$) = 8.76, $p < 0.01$ and the age × MD interaction, χ^2 (1, $N = 193$) = 4.75; $p < 0.05$. No significant main effect emerged for severity of MD, χ^2 (1, $N = 193$) = 0.10; $p > 0.05$. As shown, older children ($M = -0.57$) had more severe deficits than younger children ($M = -0.46$), but these age effects were primarily isolated to severe ($M = -0.62$ vs. −0.44 for older and younger, respectively) than the moderate sample (−0.53 vs. −0.54). These results, however, should be interpreted with caution because the analysis did not control for differences in intelligence and reading skill. Thus, a further analysis was necessary to determine whether the magnitude of the ESs on cognitive measures were a function of age and degree of MD.

Multi Level Mixed Modeling

In the final analysis, we studied whether ESs varied across age, IQ, math level, and reading level and the type of cognitive measure. We used a HLM where Level 1 equations represented the ESs when comparing MD vs. average achievers. Level 2 reflected study effects.

As shown in Table 6, the unconditional model yielded parameter estimates for the fixed effects (the intercept) for the average ES in the sample of studies. For an unconditional model, there was only one fixed effect that provided an estimate. The estimate average ES across studies was −0.47. Also shown in Table 6, both the random effects for intercept and the residual were different from zero. These estimates indicated that the studies differed significantly in their ESs. Further, there was also substantial variation (according to the size of the estimate of the residual) within the studies. For the unconditional model, we computed an interclass correlation by taking the ratio of the variance component between studies (0.07) to the sum of the variance between and within ESs ($0.07 + 3.18 = 3.25$). The interclass correlation tells us the total proportion of variance across each

Table 6. HLM Regression Predicting Effect Sizes for all Cognitive Measures Comparing Math Disabled and Average Achievers.

Unconditional model				
Fixed effect				
	Estimate	SE	*t*-ratio	*p* value
Intercept	−0.47	0.07	−6.09	<0.001
Random effect (covariance parameter estimates)				
	Estimate	SE	*Z*	*p* value
Intercept[a]	0.07	0.09	1.86	0.03
Residual[b]	3.18	0.34	9.26	<0.0001

[a] Variance between studies.
[b] Variance within studies.

individual study. The intraclass correlation was 0.02 (0.07/3.25). Thus, only 2% of the variance in ESs was at the study level whereas 98% of the variance was at the within study level.

This unconditional model provided a baseline to compare our first conditional model (conditional model) that included main effects for age, IQ and math level, ESs for IQ and Math, and reading level. The question of interest related to this conditional modeling was whether any of the classification measures, when partialed for the influence of other classification variables, would predict ESs. Table 7 shows a conditional model that entered the fixed effects for age, IQ, reading score, math level for the MD participants as well as the ES for IQ and math. The estimates for each variable shown in Table 7 have been partialed for the influence of all other variables. Because of the number of estimates, we set alpha to a conservative 0.003 in the conditional models (Note. The last conditional model shown in Table 8 has 13 variables and a Bonferroni correction yields 0.003, 0.05/13 = 0.0038). As shown in Table 7, no variable contributed significant variance in predicting the overall ESs. When comparing Tables 6 and 7, the variance component representing the difference between the studies in the conditional model changed only slightly relative to the unconditional model (a variance of 0.07 changed to 0.03). However, the within variance was reduced by 12% [(3.18−2.81)/3.18]. The results show that unique variance in IQ, math ability, and reading ability did not significantly moderate the magnitude of the ESs.

Table 7. Conditional Model Predicting Effect Sizes with Classification Measures Comparing Math Disabled and Average Achievers.

Conditional Model				
Fixed effect				
	Estimate	SE	*t*-ratio	*p* value*
Intercept	−0.58	0.21	−2.70	0.01
Age MD	−0.0001	0.002	−0.29	0.77
Classification variables				
IQ MD	0.02	0.01	1.37	0.17
ES IQ	−0.40	0.32	−1.28	0.20
Math Level	−0.006	0.03	−0.20	0.84
ES Math	0.10	0.08	1.25	0.21
Read Level	−0.0001	0.02	−0.01	0.99
Random effect (covariance parameter estimates)				
	Estimate	SE	*Z*	*p* value
Intercept	0.03	0.04	1.14	–
Residual	2.81	0.37	7.43	<0.0001

Note: Read MD = reading standard score for MD group; ES = effect size between MD and average achievers, IQ MD = Intelligence standard score for MD, Math MD = Math standard score for MD group.
*$p<0.01$.

We next tested whether the type of cognitive measure contributed unique variance. For this conditional, we coded the cognitive measures of WM, STM, LTM, speed, and problem solving as dichotomous variables (present as 1 vs. absent as 0). These measures reflected a point biserial correlation with the overall ESs. That is, the cognitive variables represented the presence of the measures (coded as 1) when compared to all other measures (coded as 0). The results are shown in Table 8. As shown the only significant effect that met the alpha level was verbal WM. The contribution of this domains was significantly better than chance. No other domain or classification variable was related to the overall ES. These findings were interpreted as suggesting that verbal WM was a major determinant of the differences between MD and NMD children. The second conditional model also eliminated the between study variance as well as substantially reduced within study variance by 20% [(3.18−2.53)/3.18]. The results also show that magnitude of the ES varied substantially across the three models (−0.47, −0.58, −0.10, respectively). The final conditional model showed

Table 8. Conditional Model Predicting Effect Sizes for all Cognitive Measures Comparing Math Disabled and Average Achievers.

Conditional Model

Fixed effect

	Estimate	SE	*t*-ratio	*p* value
Intercept	−0.10	0.18	−0.56	0.58
Age MD	−0.001	0.001	−0.19	0.85
Classification variables				
IQ MD	0.019	0.01	1.72	0.08
ES IQ	−0.15	0.27	−0.57	0.56
Math level	−0.011	0.03	−0.36	0.72
ES Math	0.10	0.059	1.75	0.08
Read level	0.01	0.01	1.16	0.24
Domain				
Problem solving	−0.41	0.15	−2.63	0.009
Speed	−0.39	0.17	−2.27	0.02
STM-words	−0.34	0.19	−1.74	0.08
STM-digits	0.24	0.20	1.18	0.24
LTM	−0.42	0.20	−2.05	0.04
Verbal WM	−0.51	0.14	−3.63	0.0004*
Visual-spatial WM	−0.40	0.19	−2.09	0.03

Random effect (covariance parameter estimates)

	Estimate	SE	*Z*	*p* value
Intercept	–	–	–	–
Residual	2.53	0.31	8.12	<0.0001*

Note: ES = effect size between MD and average achievers, IQ MD = Intelligence standard score for MD, Math MD = Math standard score for MD group.
$^{*}p<0.001$.

that magnitude of the ESs were reduced substantially when the influence of the classification and the type of cognitive variables were partialed from the analysis.

DISCUSSION

This synthesis had three purposes. First, we sought to determine whether the cognitive deficits in children with MD were distinct from their average achieving counterparts, as well as from children with RD and comorbid disorders (combined reading and MD). The results clearly indicate that moderate (0.50–0.80) weighted ESs in favor of age-matched average achieving children emerged on measures of verbal-problem solving ($M = -0.58$), naming speed ($M = -0.70$), and verbal ($M = -0.70$) and visual-spatial WM ($M = -0.63$) and LTM ($M = -0.72$). Children with MD were also differentiated from children with combined reading and MDs. ESs in favor of the MD group emerged on measures of literacy ($M = 0.75$), visual-spatial problem solving ($M = 0.51$), LTM ($M = 0.44$), STM-words ($M = 0.71$), and verbal WM ($M = 0.30$). Interestingly, an advantage was found for the comorbid group on measures of naming speed ($M = -0.39$) and attention ($M = -0.57$). In contrast to comparisons with the comorbid group, children with MD could not be clearly differentiated from children with RD on several measures (M ES $= -0.10$). However, we did find weak to moderate (between 0.20 and 0.49) ESs in favor of children with RD on measures of naming speed ($M = -0.23$) and visual-spatial WM ($M = -0.30$).

Second, we sought to determine whether the cognitive deficits in children with MD were a function of age. This was done to determine whether the magnitude of differences between MD and average achieving children persisted across different age levels. The results of the HLM analysis clearly indicated that age was unrelated to the magnitude of ESs when the influence of all other classification variables were partialed out in the analysis (see Table 7). These finding emerged even when the type of domain assessed, IQ, math level, and reading level were partialed out of the analysis. Thus, the results support the notion that MD is persistent across age.

The third question addressed whether the ESs varied as a function of severity in MD and intellectual level. We found that MD interacted with age in our preliminary analyses. Further, age effects were more pronounced in the severe math group than the moderate math group. However, the age and severity of MD effect were eliminated in the HLM analysis. Thus, IQ and severity of math differences played little roles in outcomes related to the cognitive ES variables.

In general, our results are consistent with previous syntheses of the literature that have attributed MD to memory deficits (e.g., Geary, 2004). The variables that contributed most to overall cognitive functioning of MD participants relative to NMD participants was verbal WM. More specifically, we found that memory performance of MD samples was characterized as reflecting a deficit in WM, but not STM for digits. Thus, the question emerges, how can the cognitive deficits in WM in children with MD be explained?

Three possibilities are considered. First, semantic memory deficits underlie MD. Geary (1993) suggested that semantic memory may underlie many of the problems of children with MD because learning number facts are tied to representations of semantic memory. However, Landerl et al. (2004) have challenged the underlying assumptions that these children suffer semantic memory deficits. Landerl et al. argued that semantic memory has been confounded with numerical processing. They indicate that there is little evidence of nonnumerical semantic deficits in children with MD. They also argue that semantic memory for numbers is mediated by a different system than a general memory system and that number knowledge is distinct from semantic memory. We found in our conditional model when variables related to various classification measures, naming speed, and problem solving were partialed from the analysis that distinct process related to verbal WM was related to the magnitude of the ESs. No significant differences were found on STM measures for numbers. In fact, verbal WM was the only measure found to predict overall cognitive functioning. Thus, it appears that the results suggest deficits in a verbal memory system – but not necessarily for number information.

Second, processes that underlie MD are the same as those that underlie RD. No doubt, it is considered that an important correlate of MD is RD. Lewis, Hitch, and Walker (1994) estimated that 40% of the children with RDs also have a MD. Hanich, Jordan, Kaplan, and Dick (2001) also found that children with MD were superior to children with combined disabilities in areas related to language but not in areas of visual/spatial processing or manipulation of numbers. Studies by Rourke (see Rourke, 1993, for a review) found that children with MD were more likely to have difficulties in spatial and psychomotor abilities whereas children with RD tended to have more difficulties on verbal tasks. However, Shalev, Manor, and Gross-Tsur (1997) found no quantitative differences between children with RD and MD. Their findings suggest that the distinction between children with MD and RD may be related to ADHD. The present results suggest that children with RD and MD were differentiated only on measures of naming speed and visual-spatial WM. However, the magnitude of these ESs are small. Thus, we found only partial support that groups can be differentiated on measures of visual memory.

Some studies have documented that children with MD also perform poorly on very complex math tasks such as word problems and that this is not necessarily due to just a numerical deficit, but to both phonological and executive processing deficits (Swanson & Sachse-Lee, 2001). Thus, one could argue that differences between math ability groups, such as children with RD and the comorbid group become much more reliable with greater manipulations of phonological information. Phonological STM is certainly believed to be composed of rehearsal components and phonological skills that are deficient in children with MD and RD. As shown in Table 3, the two groups could not be differentiated on measures attributed to phonological memory. That is, the ESs between these two groups was 0.16 for STM-words and 0.03 for STM-digits. The difficulty with the phonological explanation, however, is that we found an advantage for RD in terms of naming speed, a measure assumed to tap phonological processing. Thus, although we do not discount the fact that RD and MD children share similar deficits in phonological processing, some disadvantages emerged for children with MD in the memory areas not attributed to phonological skills (i.e., visual-spatial WM).

Focusing on variables independent of the classification variable, Landerl et al. compared children of different subtypes and found that children with MD were normal on several tasks involving phonological STM, accessing nonverbal information, language abilities, and psychomotor abilities. These findings are similar to ours. However, they concluded that children with MD were best defined in terms of processing of specifically numerical information. They also found that children with RD performed slightly similar to controls on numerical processing tasks. MD and RD children were slower than controls in reciting number sequences, although unlike children with MD, the number naming trend in RD children disappeared once general ability was controlled for. Although several studies (e.g., see Shalev et al., 1997) along with Landerl et al. have found that MD children differ more on measures that include numerical information than other measures, we found in our meta-analysis that these tasks are comparable between the groups (e.g., ES for STM-numbers was 0.03). No doubt our findings did not tap all the basics of numerical concepts (especially numerosity, i.e., dot counting, number comparison, and subsidizing). Our results do suggest that perhaps differences between the groups may be related to basic differences in naming speed and visual-spatial WM.

Finally, WM deficits may underlie MD. Because verbal and visual/spatial WM tasks were deficient between MD and average achievers, it appears that their memory deficits may operate outside a verbal system. This finding differs from other studies suggesting that WM deficits in MD children are domain specific. For example, Siegel and Ryan (1989) found that children with MD perform poorer on WM tests related to counting and remembering digits. They did not have difficulties on nonnumerical WM tasks. A study by McLean and Hitch (1999) also suggested that children with MD do not have general WM deficits, but have specific problems with the numerical information. In contrast, Koontz and Berch (1996) tested children with and without MD on digit and letter span tasks. They found that the children with MD performed below average on both types of tasks, indicating a general WM difficulty (also see Swanson, 1993, for a similar finding). In contrast, to the other studies Temple and Sherwood (2002) found no difference between groups on any of the measures for forward and backward digit span and no correlation was found between memory and arithmetic ability. The review by Landerl et al. suggests there is no convincing evidence that WM is a causal feature of MD. Our results showed, however, support at least for a verbal WM deficit when the influence of age, IQ and reading ability, and related domain categories (e.g., STM-number information, naming speed) are partialed out. We would argue that because variables related to STM, LTM, and visual-spatial WM were partialed from the analysis, that the residual variance related to the verbal WM measures may reflect measures of controlled processes and therefore tap a general system. No doubt, this speculation will have to be tested in subsequent studies.

In summary, our analysis of the experimental research identified cognitive differences between MD and average math achievers. The most important conclusion is that MD children as a group are distinctively disadvantaged when compared to their peers who are average in math performance across a broad range of tasks. A primary problem for students with MD was their difficulty in performing on WM tasks.

ACKNOWLEDGMENT

The writing of this chapter was partially supported by a U.S. Department of Education, Cognition and Student Learning (USDE R305H020055), U.S. Department of Education, Institute of Education Sciences awarded to the first author. The authors are indebted to Georgia Doukas, Rebecca Gregg, and Crystal Howard for coding of articles and establishing inter-rater agreements. The report does not necessarily reflect the view of the U.S. Department of Education. This report is preliminary and a more extensive and updated report is currently under journal review.

REFERENCES

Badian, N. A. (1999). Persistent arithmetic, reading, or arithmetic and reading disability. *Annals of Dyslexia*, *49*, 45–70.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brookshire, B. L., Butler, I. J., Ewing-Cobbs, L., & Fletcher, J. M. (1994). Neuropsychological characteristics of children with Tourette syndrome: Evidence for a nonverbal learning disability? *Journal of Clinical and Experimental Neuropsychology*, *16*, 289–302.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. London: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). New York: Academic Press.

Fletcher, J. M. (1985). Memory for verbal and nonverbal stimuli in learning disability subgroups: Analysis by selective reminding. *Journal of Experimental Child Psychology*, *40*, 244–259.

Garnett, K., & Fleischner, J. E. (1983). Automatization and basic fact performance of normal and learning disabled children. *Learning Disability Quarterly*, *6*, 223–230.

Geary, D. C. (1993). Mathematical disabilities: Cognition, neuropsychological and genetic components. *Psychological Bulletin*, *114*, 345–362.

Geary, D. C. (2004). Math disabilities. In: H. L. Swanson, K. Harris & S. Graham (Eds), *Handbook of learning disabilities*. NY: Guildford.

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology*, *77*, 236–263.

Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with math disability. *Journal of Experimental Child Psychology*, *88*, 121–151.

Geary, D. C., Hoard, M. K., & Hamson, C. O. (1999). Numerical and arithmetical cognition: patterns of functions and deficits in children at risk for a mathematical disability. *Journal of Experimental Child Psychology*, *74*, 213–239.

Gonzalez, J. E. J., & Espinel, A. I. G. (1999). Is IQ-achievement discrepancy relevant in the definition of arithmetic learning disabilities? *Learning Disability Quarterly*, *22*, 291–301.

Gonzalez, J. E. J., & Espinel, A. I. G. (2002). Strategy choice in solving arithmetic word problems: Are there differences between students with learning disabilities, G-V poor performance and typical achievement students? *Learning Disability Quarterly*, *25*, 113–122.

Gross-Tsur, V., Manor, O., & Sha1ev, R. S. (1996). Developmental dyscalculia: Prevalence and demographic features. *Developmental Medicine and Child Neurology*, *38*, 25–33.

Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning difficulties. *Journal of Educational Psychology*, *93*, 615–626.

Hecht, S. A., Torgessen, J. K., Wagner, R., & Rashotte, C. (2001). The relationship between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study of second to fifth grades. *Journal of Experimental Child Psychology*, *79*, 192–227.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hitch, G. J., & McAuley, E. (1991). Working memory in children with specific arithmetical learning disabilities. *British Journal of Psychology*, *82*, 375–386.

Jordan, N., Hanich, L. B., & Kaplan, D. (2003a). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development*, *74*, 834–850.

Jordan, N., Hanich, L. B., & Kaplan, D. (2003b). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, *85*, 103–119.

Jordan, N., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology*, *94*, 586–597.

Jordan, N., & Montani, T. (1997). Cognitive arithmetic and problem solving: A comparison of children with specific and general mathematics difficulties. *Journal of Learning Disabilities*, *30*, 624–634.

Klorman, R., Thatcher, J. E., Shaywitz, S. E., Fletcher, J. M., Marchione, K. E., Holahan, J. M., Stuebing, K. K., & Shaywitz, B. A. (2002). Effects of event probability and sequence on children with attention-deficit/hyperactivity, reading, and math disorder. *Biological Psychiatry*, *52*, 795–804.

Koontz, K. L., & Berch, D. B. (1996). Identifying simple numerical stimuli: Processing inefficiencies exhibited by arithmetic learning disabled children. *Mathematical Cognition*, *2*(1), 1–23.

Landerl, K., Bevan, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8–9 year old students. *Cognition*, *93*, 99–125.

Lennox, C., & Siegel, L. S. (1993). Visual and phonological spelling errors in subtypes of children with learning disabilities. *Applied Psycholinguistics*, *14*, 473–488.

Lewis, C., Hitch, G., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- and 10-year old boys and girls. *Journal of Child Psychology and Psychiatry*, *35*, 283–292.

Lindsay, R. L., Tomazic, T., Levine, M. D., & Accardo, P. J. (2001). Attentional function as measured by a continuous performance task in children with dyscalculia. *Journal of Developmental & Behavioral Pediatrics*, *22*, 287–293.

Loveland, K. A., Fletcher, J. M., & Bailey, V. (1990). Verbal and nonverbal communication of events in learning-disability subtypes. *Journal of Clinical and Experimental Neuropsychology*, *12*, 433–447.

Lucangeli, D., Coi, G., & Bosco, P. (1997). Metacognitive awareness in good and poor math problem solvers. *Learning Disabilities Research & Practice*, *12*(4), 209–212.

Lund, A. M., Hall, J. W., Wilson, K. P., & Humphreys, M. S. (1983). Frequency judgment accuracy as a function of age and school achievement (Learning disabled versus non-learning-disabled) patterns. *Journal of Experimental Child Psychology*, *35*, 236–247.

Mattson, A. J., Sheer, D. E., & Fletcher, J. M. (1992). Electrophysiological evidence of lateralized disturbances in children with learning disabilities. *Journal of Clinical and Experimental Neuropsychology*, *14*(5), 707–716.

Mazzocco, M. M. (2001). Math learning disability and math LD subtypes: Evidence from studies of Turner syndrome, Fragile X syndrome, and Neurofibromatosis type 1. *Journal of Learning Disabilities*, *34*, 520–533.

McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetical difficulties. *Journal of Experimental Child Psychology*, *74*, 240–260.

Miles, J., & Stelmack, R. M. (1994). Learning disability subtypes and the effects of auditory and visual priming on visual event-related potentials to words. *Journal of Clinical and Experimental Neuropsychology*, *16*, 43–64.

Montague, M., & Applegate, B. (1993). Mathematical problem-solving characteristics of middle school students with learning disabilities. *The Journal of Special Education*, *27*, 175–201.

Nolan, D. R., Hammeke, T. A., & Barkley, R. A. (1983). A comparison of the patterns of the neuropsychological performance in two groups of learning disabled children. *Journal of Clinical Child Psychology*, *12*, 22–27.

Passolunghi, M. C., Cornoldi, C., & De Liberto, S. (1999). Working memory and intrusions of irrelevant information in a group of specific poor problem solvers. *Memory & Cognition*, *27*, 779–790.

Passolunghi, M. C., & Siegel, L. S. (2001). Short-term memory, working memory, and inhibitory control in children with difficulties in arithmetic problem solving. *Journal of Experimental Child Psychology*, *80*, 44–57.

Rourke, B. P. (1993). Arithmetic disabilities, specific and otherwise: A neuropsychological perspective. *Journal of Learning Disabilities*, *26*, 214–226.

SAS Institute Inc. (1999). *SAS/STAT user's guide, version 7*. Cary, NC: SAS Institute Inc.

Shafrir, U., & Siegel, L. S. (1994). Subtypes of learning disabilities in adolescents and adults. *Journal of Learning Disabilities*, *27*, 123–134.

Shalev, R. S., Manor, O., & Gross-Tsur, V. (1997). Neuropsychological aspects of developmental dyscalculia. *Mathematical Cognition*, *3*(2), 105–120.

Share, D. L., Moffitt, T. E., & Silva, P. A. (1988). Factors associated with arithmetic-and-reading disability and specific arithmetic disability. *Journal of Learning disabilities*, *21*, 313–320.

Siegel, L. S., & Ryan, E. B. (1988). Development of grammatical-sensitivity, phonological, and short-term memory skills in normally achieving and learning disabled children. *Developmental Psychology*, *24*(1), 28–37.

Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development*, *60*, 973–980.

Sikora, D. M., Haley, P., Edwards, J., & Butler, R. W. (2002). Tower of London test performance in children with poor arithmetic skills. *Developmental Neuropsychology*, *21*, 243–254.

Snijders, T. A., & Bosker, R. J. (2003). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Swanson, H. L. (1993). Working memory in learning disability subgroups. *Journal of Experimental Child Psychology*, *56*, 87–114.

Swanson, H. L. (1994). The role of working memory and dynamic assessment in the classification of children with learning disabilities. *Learning Disabilities Research & Practice*, *9*(4), 190–202.

Swanson, H. L., Cooney, J. B., & Brock, S. (1993). The influence of working memory and classification ability on children's word problem solution. *Journal of Experimental Child Psychology*, *55*, 374–395.

Swanson, H. L., & Rhine, B. (1985). Strategy transformations in learning disabled children's math performance: Clues to development of expertise. *Journal of Learning Disabilities*, *18*, 409–418.

Swanson, H. L., & Sachse-Lee, C. (2001). Mathematical problem solving and working memory in children with learning disabilities. Both executive and phonological processes are important. *Journal of Experimental Child Psychology*, *79*, 294–321.

Temple, C., & Sherwood, S. (2002). Representation and retrieval of arithmetical facts: Developmental difficulties. *Quarterly Journal of Experimental Psychology*, *55A*(3), 733–752.

APPENDIX. SAMPLE STUDIES EXCLUDED FROM META-ANALYSIS

Reason for Exclusion

Author(s) and Publication year

No IQ scores ascribed to specifically to children with MD

- Geary (1990)
- Geary and Brown (1991)
- Geary, Brown, and Samaranayake (1991)
- Wilson and Swanson (2001)
- Barwick and Siegel (1996)
- Jordan, Huttenlocher, and Levine (1992)
- Jordan, Kaplan and Hanich (2002)
- Hanich et al. (2001)
- Gallagher, De Lisi, Holst, McGillicuddy-De Lisi, and Morely (2000)
- Swanson (1993)
- Jordan and Hanich (2000)
- Frensch and Geary (1993)
- Fuchs, Fuchs, Hamlett, and Stecker (1991)
- Fuchs, Fuchs, and Karns (2001)
- Bottge, Heinrichs, Chan, and Serlin (2001)
- Bottge (1999)
- Geary, Bow-Thomas, and Yao (1992)
- Siegel and Ryan (1989)
- Jordan and Montani (1997)

No standardized math tests given

- Fuchs, Fuchs, Phillips, Hamlett, and Karns (1995)
- Shalev, Weirtman, and Amir (1988)
- Parmar, Cawley, and Frazita (1996)

No standardized IQ or math measure reported for the targeted sample

- Kosc (1974)
- Jordan, Levine, and Huttenlocher (1995)
- Shalev, Manor, Amir, and Gross-Tsur (1993)
- Jordan, Levine, and Huttenlocher (1994)
- Levine, Jordan, and Huttenlocher (1992)

Insufficient information to calculate ES

- Kosc (1974)
- Robinson, Menchetti, and Torgesen (2002)
- Jordan (1995)
- Bryant, Bryant, and Hammill (2000)
- Knopik (1996)
- Montague (1997)
- Rourke (1993)
- Ansari and Karmiloff-Smith (2002)
- Levy (1981)
- Siegel (1974)
- Tishler (1981)
- Shalev, Auerbach, and Gross-Tsur (1995)

MD sample scores could not be separated from total sample

- D'Amato, Dean, and Rhodes (1998)
- Swanson and Cooney (1985)
- Shafrir, Ogilvie, and Bryson (1990)
- Vargo, Grosser, and Spafford (1995)
- Farrag, Shaker, Hamdy, and Wafaa (1995)
- Zentall, Smith, Lee, and Wieczorek (1994)
- Matochnik, Rumsey, Zametkin, Hamburger, and Cohen (1996)
- Janssen, Boeck, Viaene, and Vallaeys (1999)
- Feagans and Appelbaum (1986)
- Marshall, Schafer, O'Donnell, Elliott, and Handwerk (1999)
- Morris et al. (1998)
- Fuchs et al. (1998)
- Fletcher et al. (1994)
- Fuchs, Fuchs, Karns, Hamlett, and Katzaroff (1999)
- Geary and Burlingham-Dubree (1989)
- Fuchs, Fuchs, Hamlett, and Karns (1998)
- Fuchs, Fuchs, Hamlett, Phillips, Karns, and Dutka (1997)

No average achieving comparison group

- Cirino, Morris, and Morris (2002)
- Buono et al. (1998)
- Naglieri and Johnson (2000)
- Ozols and Rourke (1988)
- Gross-Tsur et al. (1996)

- Fuchs, Fuchs, Hamlett, and Appelton (2002)
- Shalev, Manor, Amir, Wertman-Elad, and Gross-Tsur (1993)
- Fuchs and Fuchs (2002)
- Davis, Parr, and Lan (1997)
- Strang and Rourke (1983)
- Rosenberg (1989)
- Ackerman and Dykman (1995)
- Goldstein, Katz, Slomka, and Kelly (1992)

SUMMARIZING QUALITATIVE RESEARCH IN SPECIAL EDUCATION: PURPOSES AND PROCEDURES

Thomas E. Scruggs, Margo A. Mastropieri and Kimberly A. McDuffie

ABSTRACT

In recent years, there has been an extraordinary accumulation of qualitative research in special education. However, as yet, there has been little accumulation of the understandings gained from these studies. This omission has important implications for knowledge development, the utilization of findings in practice, and providing implications for policy. In this chapter, we review and discuss perspectives and procedures from other fields with respect to aggregation of qualitative data. Additionally, we propose a specific method for the meta-synthesis of qualitative research in the area of special education. This synthesis would not be a numerical compilation of outcomes, as in traditional meta-analysis, but would treat individual research reports as "informants," and employ procedures, such as analytic induction and the constant comparative method to develop higher understandings across individual cases. Such efforts are thought to be essential to reaching higher analytic goals and also to enhancing the

Applications of Research Methodology
Advances in Learning and Behavioral Disabilities, Volume 19, 315–335

ISSN: 0735-004X/doi:10.1016/S0735-004X(06)19012-6

generalizability of qualitative research. It is argued that meta-synthesis efforts could do much to promote the impact of the shared understandings gained from individual qualitative research efforts.

In recent years, there has been an explosion in the amount of qualitative research conducted in special education. For example, a search of Dissertation Abstracts, using the descriptors "special education" and "qualitative," reveals enormous growth in special education dissertations using qualitative methodology. From a single qualitative dissertation identified in 1979, the number and proportion of qualitative dissertations have grown steadily, until today, with approximately 120 per year, or nearly 20% of the total.

This extraordinary rate of growth is depicted graphically in Fig. 1. Evaluation of databases cataloging ERIC documents and relevant research

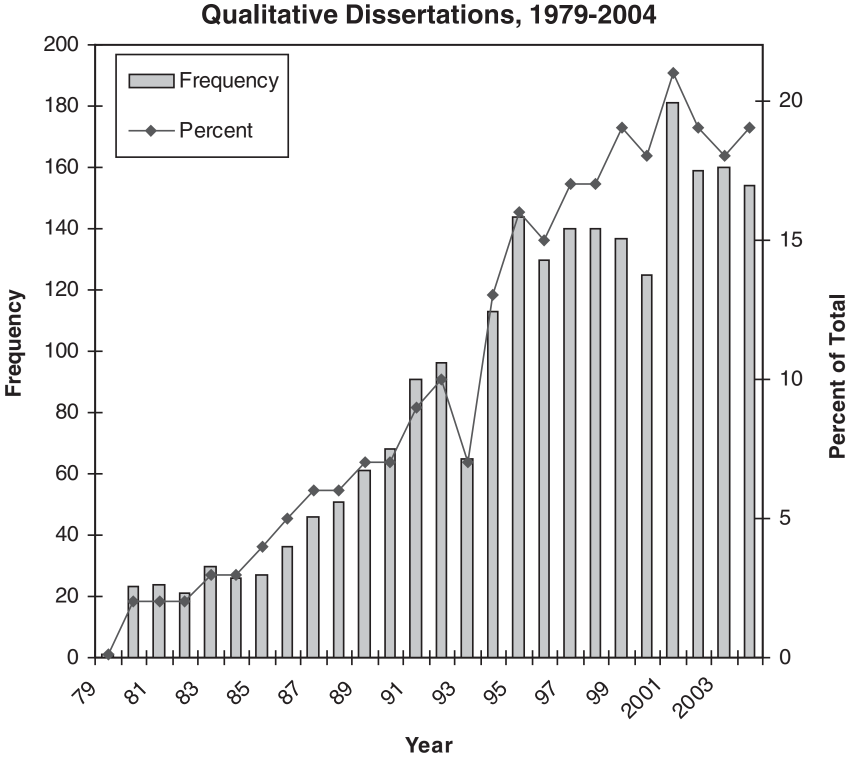

Fig. 1. Frequency and Percent of Qualitative Dissertations, 1979–2004.

journals (*ERIC*, *PsycInfo*) and hand searches of special education journals reveal similar, although subtler, trends. Although an analysis using more specific and refined descriptors may reveal different numbers, it seems clear that the sheer number of qualitative research being conducted in special education today is very considerable.

Although the reasons for this increase in volume are not entirely known, it seems likely that contributing factors include: (a) the limitations of exclusively quantitative methods to explain satisfactorily all aspects of special education; (b) the relative strengths of qualitative methods in providing authentic, rich descriptions, and promoting carefully considered understandings of special education; (c) an increasing willingness of editors and dissertation committees to accept qualitative methodology; and (d) an increasing number of scholars with skills in qualitative methodology (Scruggs & Mastropieri, 1995). Regardless of reason, it is clear that a substantial amount of qualitative research has been conducted, with the potential of greatly developing understandings of special education practice. Qualitative researchers have provided important insights relevant to enhanced understanding of a wide variety of areas relevant to special education, including, for example, paraeducator experiences (Downing, Ryndak, & Clark, 2000; Marks, Schrader, & Levine, 1999), narratives of Latino mothers of children with disabilities (Skinner, Bailey, Correa, & Rodriguez, 1999), transition practices (Collet-Klingenberg, 1998), special education teacher roles (Weiss & Lloyd, 2002), and inclusive schooling (Kozleski & Jackson, 1993; Salisbury, Palombaro, & Hollowood, 1993; Scruggs & Mastropieri, 1994).

However, at present, only a minimal impact of qualitative studies has been realized. In sharp contrast to quantitative research, there has been little accumulation of understandings gained from these studies. One reason for this is that procedures for synthesizing special education research employing other methodologies have been developed and implemented (e.g., Scruggs & Mastropieri, 1996, 2000, 2001). At present, the accumulated findings of qualitative research have had little impact, compared with the very substantial potential of such research to elevate understandings. Years ago, Yin and Heald (1975) argued that while "each case study may provide rich insights into a specific situation, it is difficult to generalize about the studies as a whole" (p. 371). Valuable individual research findings have not been situated in a larger context, and have not been presented in a usable form necessary for the real world of practice and policy making. Perhaps in part because of this fact, some influential policymakers have suggested that qualitative research may not be appropriate for drawing policy implications.

And in fact, at present, there has been only limited impact of qualitative research on daily practice or shared understandings of special education.

In order to elevate and promote accumulated understandings from qualitative research, procedures are needed to summarize and synthesize findings of qualitative research, to "find ways to aggregate, compare, or contrast already existing studies" (Schofield, 1990, p. 222). Such "meta-synthesis" refers not to secondary analyses of data created from data pooled from individual qualitative research studies (e.g., West & Oldfather, 1995), but rather to "theories, grand narratives, generalizations, or interpretive translations produced from the integration or comparison of findings from qualitative studies" (Sandelowski, Docherty, & Emden, 1997, p. 366). Qualitative meta-synthesis is largely unknown in education, but has been proposed and implemented in other fields, such as the health sciences. Such efforts could enhance the present limited visibility and impact of qualitative research, directly address commonly cited problems with across-case generalization of findings, and improve implications for policy and practice. Considered separately, individual studies have limited accessibility and even less impact. However, carefully conducted synthesis efforts could greatly improve visibility and allow multiple voices of individual researchers to be heard as a community.

CONCERNS ABOUT META-SYNTHESIS

Several important objections can be raised to the synthesis of qualitative research on theoretical as well as on practical grounds. These arguments have been summarized by Sandelowski et al. (1997). Not the least important is the fact that qualitative research, by its nature, seems antithetical to synthesis efforts, and in fact may be endangered by this process. Because of the complexities inherent in the in-depth study of individual cases, qualitative studies seem to resist "summing up" (Light & Pillemer, 1984). The idiographic nature of qualitative research seems to argue against synthesis, in that the uniqueness of individual projects could be lost; further, such synthesis could represent a departure from the larger pedagogic and emancipatory aims of some qualitative research. In fact, it can be argued that it is precisely this idiographic element that provides such a sharp contrast with quantitative studies, which provide general conclusions about groups and are less relevant to individual cases.

When qualitative research contains arguably aesthetic components, containing elements shared with novels in describing elements of human experience, summarization can be particularly questionable. Sandelowski et al. (1997, p. 366) asked, rhetorically, "Can you sum up a poem?"

Another source of difficulty in synthesizing qualitative research is the diversity of the qualitative research that exists. In fact, it has been argued that collecting research including a variety of methodologies and perspectives – including, for example, case studies, phenomenological studies, ethnographies, semi-structured interviews, and narratives – under a general umbrella of "qualitative research" is misleading, and may trivialize significant differences among them (Atkinson, 1995). Synthesis of such a wide variety of perspectives and methods could be particularly problematic.

A final concern for meta-synthesis of qualitative research is that criteria for evaluating study quality, like criteria for many areas of human endeavor, are context-dependent, and may not be consistent across individual research efforts. Lincoln (1995) described criteria for establishing quality of qualitative research as "emerging," while others have argued against establishment of uniform standards of quality as "criteriology" (Schwandt, 1996). Nevertheless, some standards have been recently voiced for evaluating qualitative research in special education (Brantlinger, Jiminez, Klingner, Pugach, & Richardson, 2005).

In summary, a number of concerns can be expressed about synthesis efforts for qualitative research. Not the least of these is the emphasis on "$N = 1$ experiences" (Eisner, 1991, p. 197) that argues against "adding up" these experiences.

THE COST OF *NOT* SYNTHESIZING QUALITATIVE RESEARCH

Although a number of concerns have been expressed about qualitative meta-synthesis, there are also dangers associated with *not* summarizing qualitative research. One important concern that has been expressed is that qualitative researchers have been isolated from each other and work in a "cottage industry," producing "one shot research" (Estabrooks, Field, & Morse, 1994, p. 510). As a consequence, researchers have little opportunity to learn from each other, and are in a position of continually reinventing the wheel. In special education, qualitative researchers too often have failed to situate their work in larger programs of research or scholarship (Scruggs & Mastropieri, 1995).

Years ago, Glaser and Strauss (1971, p. 181) warned that the failure to develop local grounded theories into more formal, generalizable theories would relegate individual findings into "little islands of knowledge," unconnected with one another and unvisited by other researchers. This is not a trivial concern; without developing the connectedness latent within and across qualitative research studies, this important body of research may exert only minimal and tentative impact on the field of special education, and becomes more vulnerable to charges that qualitative research should not be employed in influencing policy decisions.

ADVANTAGES OF QUALITATIVE META-SYNTHESIS

Qualitative meta-synthesis could represent one method of developing understandings across individual studies. Sandelowski et al. (1997) have argued that efforts to synthesize qualitative research studies are essential to reaching higher analytic goals. They referred to the process as "analytic interruptus" (p. 366), when qualitative researchers fail to go far enough in their work; that is, when they fail to reveal the connections among findings.

Qualitative research has been seen by some as "ungeneralizable," but this is true only when generalization is narrowly and incorrectly conceived solely in terms of sampling and statistical significance. Since qualitative research is "directed toward the kind of generalizations about particulars," it is "indefensible, dysfunctional, and out of touch with contemporary views of science not to recognize and value these generalizations" (Sandelowski et al., 1997, p. 176).

It has been suggested that reducing findings to a common metric would necessarily destroy much of the integrity and vitality of individual findings. However, qualitative meta-synthesis is not about summing up, averaging, or otherwise reducing findings to a "common metric." Rather, qualitative meta-synthesis can serve to enlarge the interpretive possibilities of findings and can construct larger narratives or general theories (Sandelowski et al., 1997).

Qualitative research can be synthesized in much the same way as original qualitative research is conducted. This parallel is based in part on the fact that qualitative research frequently considers data from multiple cases relevant to the purpose of the research. For example, consider a qualitative study of inclusive practices in an individual school. In this case, individual teachers, students, parents, and administrators may be interviewed, school practices can be observed, and other relevant data sources can be examined.

These data are collected and analyzed across cases, and larger generalizations and general themes are developed. Unfortunately, the degree to which emerging themes from the individual study reflect the community of studies of inclusive schools may not be adequately addressed. In a qualitative meta-synthesis effort, the voices of a number of individual researchers who have each studied inclusive schools can be treated as cases (see, for example, Anzul, Evans, King, & Tellier-Robinson, 2001). These voices can be evaluated for emerging themes and generalizations in a higher-level analytic inductive process. Generalizations, thus gained, could supply important information about educational inclusion.

PROPOSED MODELS FOR SYNTHESIZING QUALITATIVE RESEARCH

Synthesis of qualitative research is not a new idea. In the general education research literature, several proposals for synthesizing qualitative research have been made, and some initial efforts have been completed.

Noblit and Hare (1988) explored how one might pursue a meta-synthesis of published field studies. Calling this kind of synthesis "meta-ethnography," they described at least three ways by which a meta-ethnography could be constructed. One method of synthesis is "reciprocal translation," in which each study that is read and analyzed helps inform the next study as well as provides a reference for reanalyzing previously read studies. If studies investigate similar topics, then one can search for the themes or metaphors that researchers have chosen to explain what is taking place in what they have studied. It then becomes a matter of determining which metaphors from each study are the most salient, and most aptly convey common understandings.

The second method of synthesis is used when research studies about similar things come to different conclusions. A famous example of a refutational ethnography is Freeman's (2000) refutation of anthropologist Margaret Mead's findings on Samoan culture. Refutational studies may also be synthesized, and they can reveal how ideas affect interpretations. A third kind of ethnography is termed a line-of-argument synthesis. In this kind of synthesis, studies are translated into one another, but the outcome is a more parsimonious but encompassing understanding of the phenomenon being studied. The whole is greater than the parts, and it is an enlarged understanding of disparate research findings that is achieved through the line-of-argument synthesis.

Schofield (1990), in the context of increasing generalizability in qualitative research, conceived of qualitative meta-synthesis as the creation of cross-case generalizations based upon generalizations made from, and about, individual cases. Schofield suggested ways in which qualitative researchers might design their studies from the outset to make them more generalizable.

Creating a set of highly structured questions can be used to cull from qualitative data answers that can be transformed into data amenable to statistical analysis. Miles and Huberman (1994) have suggested this procedure for aggregating data from multisite studies as potentially useful in some cases. Miles and Huberman make a helpful distinction between variable-oriented and case-oriented analyses. In the latter analysis, attention is directed toward the confluence of variables or the flow of events within each case and then across cases. Yin and Heald (1975), the originators of the method, list some limitations of this method, which are: (a) if the number of cases is small, statistical techniques may lack power; (b) the number of variables worthy of coding may be large compared to the number of cases; (c) unique factors that may be critical to understand certain cases are ignored; and (d) the focus is on outcomes rather than the process.

Ragin (1987) described a ‘‘qualitative comparative’’ approach, which employs Boolean algebra, the algebra of sets and logic. A holistic view of individual cases is maintained by this method, which is not dependent on statistical analysis. What is required are data that allow one to build ‘‘truth tables,’’ i.e., categorical information on the major variables of most importance to the analysis. Ragin (1987) argues that this approach allows the investigator to examine complex and multiple patterns of causation, to produce direct and parsimonious explanations, to study cases both as whole and as parts, and to evaluate competing explanations (see also Sandelowski et al., 1997; Schofield, 1990).

PREVIOUS META-SYNTHESIS APPROACHES

To date, several meta-synthesis projects have been completed using different methods. Gersten and Baker (2000) conducted a ‘‘multi-vocal synthesis,’’ a procedure proposed for use with topics for which there is diverse writing but little systematic research (Ogawa & Malen, 1991). These authors examined the area of instructional techniques for English-language learners. Their synthesis included intervention studies using experimental designs; descriptive studies of instructional practices; and an uncommon third source, input from professional work groups. They followed methods of analysis recommended by qualitative researchers, such as Noblit and Hare (1988) and Miles and Huberman (1994), and also included five professional work groups in five different states sequentially refining a set of principles and practices that begin to define the best practice in this area. The major principles applied to the multivocal synthesis included the following (Gersten & Baker, 2000, pp. 43–44):

1. input from practitioners for generating and refining interpretations (Ogawa & Malen, 1991);
2. *triangulation* across various data sources;
3. use of *propositions* from environment and published research to provide guidance and direction;
4. use of the *constant-comparative* method across data sources to develop and refine interpretations;
5. explicit consideration of rival hypotheses (Noblit & Hare, 1988); and
6. reciprocal translation (Noblit & Hare, 1988).

Gersten and Baker conducted this investigation with eight quantitative research studies and 15 studies that analyzed classroom instruction. Input from professional work groups was conducted concurrently with analysis of research, and each influenced the other.

Braden (1992) conducted a research synthesis on hearing impairment and intelligence. Braden put written descriptions of intelligence into categories so that the data could be added to the data of quantitative studies. Although this is an example of a type of qualitative research synthesis, in the present perspective, it represents more of an attempt to quantify elements of qualitative studies and thus include them with quantitative research.

Beck (2001) synthesized 14 qualitative studies of caring within nursing education, employing Noblit and Hare's (1988) model. Research studies were divided into caring among nursing faculty and caring between faculty and students. Within each category, Beck produced a list of components and effects of caring. Five overall metaphors or themes permeating caring in nursing education emerged.

Jensen and Allen (1994) conducted a meta-synthesis of 112 qualitative research studies on health, disease, wellness, and illness. Noblit and Hare's (1988) model of reciprocal translation was employed for the analysis. The purpose of the research synthesis was "to derive substantive interpretations about health, disease, wellness, and illness from grounded theory, phenomenological, and ethnographic perspectives" (p. 349). The research reports were analyzed and compared, and new interpretations were created, based upon a synthesis of reciprocal translation. Jensen and Allen grouped studies according to research design and synthesized within those design groups. In conclusion, they constructed an overall theory of health, disease, wellness, and illness.

Campbell et al. (2003) synthesized seven qualitative studies on lay experiences with diabetes care. Six key concepts were identified from the seven, considered to be of importance in helping persons with diabetes to attain balance, well-being, and control. These key concepts included: time and experience, trust in one's self, taking a less subservient role with care providers, strategic non-compliance with medication, effective support from care providers, and an acknowledgement of the seriousness of diabetes. Interestingly, none of the included studies referenced any of the earlier papers; neither did they appear to have taken account of, or built upon, any of the previous findings. This finding provided some support that qualitative research presently represents a "cottage industry."

USING NVIVO SOFTWARE FOR META-SYNTHESIS

Given the complexity of synthesizing a large number of original research reports, each containing its own varied data sources, it appears that software designed for qualitative research studies could be employed to conduct a meta-synthesis of research in special education. NVivo software (described in more detail in the following section) may be particularly useful for entering text and other information, coding and categorizing qualitative data, and assisting with organization of qualitative data into general themes. Using the voices of individual researchers as informants, a large number of qualitative investigations of special education could be considered simultaneously. Data analysis in this type of investigation is inductive. Analytic induction "involves scanning the data for categories of phenomena and for relationships among such categories, developing working typologies and hypotheses upon an examination of initial cases, then modifying and refining them on the basis of subsequent cases" (LeCompte & Preissle, 1993, p. 254). Data from original research reports could be assimilated and evaluated in order to develop hypotheses about the practices and perspectives on inclusive education. Similar to qualitative data analysis of original data, discrepant cases and negative cases could be used to further understanding and refine hypothetical constructs. Observations and themes from original research can be subjected to the constant comparative method, in which incidents, categories, and constructs are subjected to overlapping comparisons (LeCompte & Preissle, 1993). Upon completion, the synthesis of a substantial number of original qualitative research reports of special education will enable researchers to reach higher analytic goals and enhance the generalizability and impact of this important body of research. While individual research reports can provide important examples and analytic understandings of inclusive education, qualitative meta-synthesis will allow for broader understandings of a wide variety of inclusive practices, and also provide implications for policymakers based upon the voices of a substantial number of researchers. In addition, these procedures can provide an initial model for the synthesis of qualitative research in a number of other important areas.

The conceptual framework underlying this research relates to the secondary analysis and synthesis of original research reports. This framework suggests that findings of individual research reports are not complete until they have been organized and refined into a set of higher-level understandings of the topics investigated. Unlike quantitative synthesis ("meta-analysis") of individual quantitative group-experimental studies, qualitative meta-synthesis is not concerned with summing up, averaging, or otherwise reducing findings to a "common metric." Rather, themes and insights gained from individual qualitative research could be integrated into a higher-order synthesis that provides for broad understandings of the entire corpus of research, while still respecting the integrity of individual studies.

PROCEDURES

Qualitative research synthesis could be undertaken in four parts: (a) identifying and obtaining original research reports, (b) studying and coding research using NVivo software, (c) synthesizing themes and understandings of the research, and (d) disseminating project findings.

Identifying and Obtaining Original Research Reports

The first step in an integrative review of research is to define and delimit the topic (Jackson, 1980). In the present instance, researchers should include for consideration any study that meets specific criteria. For a hypothetical example, in the area of educational inclusion of students with disabilities, the criteria might include the following:

1. The research will describe an in-depth study of inclusive schooling. Inclusive schooling is defined as any attempt to integrate one or more

students with disabilities into age-appropriate general education classrooms for all or most of the school day. While a particular focus would be on schools, which employ "full inclusion" models that serve all students with disabilities entirely within general education settings (Ferguson, 1996), studies of very intensive inclusion in individual classes, when the overall focus of the study is appropriate, would be included.

2. The study must involve the inclusion of one or more students meeting current federal standards of disability in general education settings. Studies would be included if they employ as participants students with, for example, mental retardation or hearing impairments, but not if they employ students defined as, for example, "at risk" or "low achieving."
3. The study must employ qualitative research methodology, meeting agreed-upon characteristics of qualitative methodology (e.g., LeCompte & Preissle, 1993).
4. The study must be an intensive investigation of inclusive schooling practices, be of no less than six months in duration, and should include a variety of data sources, such as interviews, classroom observations, and student products.
5. The study had been disseminated as a book, book chapter, journal article, dissertation or thesis, ERIC document, Final Report, or other unpublished report no earlier than 1990. This standard was selected to place research within a specific time frame, so that general understandings of word usage and variables studied would have greater consistency across studies.

With the topic defined and delimited, the next step is obtaining research reports. In order to obtain all possible reports meeting selection criteria, the following search procedures would be implemented:

1. Computer search of relevant databases, including *Dissertation Abstracts International*, *PsychInfo*, *ERIC*, *First Search*, and *Web of Science*. Descriptors would include *inclusion*, *full inclusion*, and *mainstreaming* as well as all disability categories (e.g., *emotional disturbance*, *behavioral disorders*, *autism*, *mental retardation*, *hearing impairments*, *deafness*); and methodological descriptors (e.g., *qualitative, ethnographic*, *ethnography*, *case study*, *narrative*, *interview*). A wide variety of descriptors should be employed, since different databases employ different search descriptors. Library personnel expert in computerized literature searches would also be consulted to identify the most relevant research possible.
2. Library search of books relevant to educational inclusion.
3. An ancestry search (Cooper, 1982), in which reference lists of all identified studies are examined for additional research reports.

4. A descendant search, in which *Social Sciences Citation Index* is employed to locate research reports that have cited identified studies.
5. Consultation with identified experts in educational inclusion and qualitative research methodology. These experts would be identified from information gained from literature search procedures as well as nominations by colleagues in special education.
6. Hand search of all major journals likely to publish research on educational inclusion, for example, *Exceptional Children*, *Remedial and Special Education*, *The Journal for the Association for Persons with Severe Handicaps*, *Exceptionality*, *Education and Training in Developmental Disabilities.*

When all research reports, which appear to meet criteria, are obtained, these reports would then be examined carefully to determine whether they should be included in the meta-synthesis. It is estimated from a preliminary search of relevant databases that a large number of research reports may meet inclusion criteria for such a synthesis. This number of reports would appear to be very large for accommodation within a single meta-synthesis effort; nevertheless, accommodation of a large number of studies is one purpose for conducting such a synthesis. Furthermore, there are reasons to believe it could be accomplished. Some original qualitative research has included large numbers of informants. Skinner, Bailey, Correa, and Rodriguez (1999), for example, collected narratives of 150 Latino mothers of children with disabilities, and were able to identify several common themes as well as provide detailed information on individual cases. These authors focused particularly on narratives in which "mothers talked specifically about themselves in relation to the child and constructed meanings of themselves and their lives around the event of disability" (p. 486). In addition, a meta-synthesis of qualitative research of this scale has been completed in the health professions (Jensen & Allen, 1994, described previously). If original research using up to 150 cases can be conducted successfully, it seems possible to conduct a qualitative meta-synthesis of a similar number of research reports as "cases."

Studying and Coding Research Using NVivo Software

Once research reports are identified and organized into a filing system, each report should be studied carefully. The reader should take careful notes of reflections and insights gained from each report, themes identified, and relevance and relation to other studies would be documented as readers proceed through the literature. The process would be an iterative one, in which previously read reports would be continuously revisited, as new insights are gained from reading additional reports (cf. Noblit & Hare's, 1988, "reciprocal translations"). This process of reading, study, and reflection is roughly analogous to the collection of field notes and journal entries in original qualitative research.

After the initial study of all project documents, researchers would scan and enter relevant data from original reports into the NVivo software program. NVivo software, also known as QSR NUD*IST Vivo (Fraser, 1999), was developed by Qualitative Solutions and Research Priority of Australia for use in analysis of qualitative research. There are a number of advantages in using NVivo for a meta-synthesis of qualitative research in special education:

NVivo can organize large amounts of data. A large number of qualitative studies may be analyzed and synthesized in a qualitative research synthesis. NVivo software can store as much of the full text of each qualitative research study as the researcher chooses. With large numbers of research studies, each with a large amount of textual information, it can be very useful to take advantage of the storing and sorting capabilities of computer technology.

Previous meta-syntheses of research from other fields have shown that syntheses of qualitative research for an entire field can be accomplished, but to date they have not demonstrated a way of analyzing large numbers of research findings that is efficient, nor have they demonstrated ways of incorporating new and recent qualitative research into ongoing meta-synthesis efforts. The use of this software would solve both of these problems by allowing researchers efficient storage and retrieval of large quantities of qualitative data. Once the body of qualitative research in special education has been entered into NVivo and analyzed, it is a much easier task to update research findings by incorporating new research findings as they are published. Because it processes text primarily, rather than numbers, NVivo is ideal for handling the "thick description" of qualitative research.

One of the concerns about any attempt to synthesize qualitative research findings is that in the process, qualities and insights unique to a particular research setting would be overlooked. Qualitative research is not meant to be aggregated statistically so that it can be generalized to the larger population. Care must be taken to avoid missing insights that are embedded in descriptive studies of particular settings.

One way to preserve each research study's individual character is to preserve as much of the author's own words as possible and to work directly from each study. The full results and conclusions' section of each research study or even the full-published study can be scanned into the program using an electronic scanner. When documents are retrieved online, they can easily be copied and inserted into NVivo. The original document is then coded by the researcher who is looking for comparisons and insights that may be generalized across studies. In this way, the meta-synthesis respects the value of the descriptions of each qualitative study, and it avoids reducing data too early in the synthesis process. In addition, researcher notes and journal entries can also be entered and coded for relevant themes in NVivo.

NVivo software allows the researcher to "think through" the analysis as it unfolds, while storing insights that may be progressively refined. The intellectual challenge of meta-synthesis of qualitative research is to hold all of the understandings from each piece of research in mind as more and more research is read and understood and as common themes and understandings emerge for the researcher. The human mind is capable of synthesis. However, while the mind may be capable of holding the particulars of several research studies at once, it is substantially more difficult to consider simultaneously dozens, or even hundreds of such studies. And it would be an impossibility to keep demographic data from all studies straight in order to cross them with research findings.

Using NVivo, the researcher first reads all of the research reports and notes previously collected on these reports. Then the researcher identifies initial codes for coding the data from those articles. The researcher then codes the text from those research articles, using the codes at what is called in NVivo "free nodes." The researcher does not commit at that point to the understanding of the research that is suggested by the initial codes, but it is a starting place. As coding proceeds, the researcher should get more ideas of how understandings from each of the qualitative research reports that are being examined may be synthesized into greater understanding of the topic of the research. With progressively more refined understanding of the research topic being examined, the researcher takes the meta-synthesis to the next level of refinement or abstraction by grouping the coded text into what is known as "tree nodes." Upon reflection of the data that is grouped by "tree nodes," the researcher can group "tree nodes" into successive levels of abstraction of "sets." The software allows the researcher to refine understanding across research reports without coming to conclusions too early in the analysis.

Conclusions about learning across studies can unfold successively gradually, and the computer can hold the researcher's initial and tentative understandings, while more defined understanding of the research emerges with greater familiarity and greater insight into the qualitative research. The software allows the researcher to cross learnings from research on certain practices in special education with demographic information about students. For example, certain special education practices might produce positive results for middle-class children but be shown to be less productive for students from lower socioeconomic backgrounds, or vice versa. Or, a practice may appear effective with certain age groups but not effective (or not studied) with other age groups. Or, a practice or perspective may appear more relevant for one specific type of disability than others. Use of the software enables the researcher to make very fine distinctions when comparing special education research results. NVivo is widely used by qualitative researchers, and successful use of this software can promote use of this technology for meta-synthesis of research in special education in future and in other fields.

Qualitative research synthesis is a valuable partner to traditional quantitative research as it illuminates the application of special education practices in particular education settings. Because knowledge gained from qualitative research is not now factored into syntheses of research in special education, much that is valuable is lost to policy-makers and others attempting the grasp of the present state of knowledge in special education.

This particular software, NVivo, is widely enough used by qualitative researchers that its application to research synthesis would allow many researchers to examine the body of qualitative research for the light that it has to shed on success in special education. Meta-synthesis of qualitative research now seems a daunting undertaking, completed by a pioneering few with the time necessary to complete the thorough, detailed analysis that the task demands. Successful use of qualitative software for this purpose would advance the inclusion of valuable insights of qualitative research into the body of knowledge in special education. It would also represent a new model to provide for other researchers, who are attempting meta-synthesis of qualitative research in other areas relevant to special education.

Synthesizing Themes and Understandings of the Research

As research reports are studied and scanned and entered into NVivo, a process of analytic induction would be employed, in which the data are scanned to identify categories and to identify relationships among these categories. Working typologies and hypotheses would be developed based upon an examination of initial cases, and then by modifying and refining these hypotheses on the basis of subsequent cases (LeCompte & Preissle, 1993). Data from the original research reports would be assimilated and evaluated in order to develop working hypotheses about the phenomena under study. Discrepant and negative cases are used to further general understandings of inclusive education and to refine hypothetical constructs. That is, if individual instances of specific challenges or problems associated with inclusive education are identified in individual reports that are at variance with the general themes from the body of research studies, specific evaluation is undertaken to determine the reasons for particular discrepancies. Resolution of such discrepant cases, and their understanding within a larger framework, can greatly enhance the understanding of inclusive practices. Observations and themes from original research, as they are developed, would be subjected to the constant comparative method, in which incidents, categories, and constructs are subjected to overlapping comparisons (LeCompte & Preissle, 1993).

AN EXAMPLE: META-SYNTHESIS OF CO-TEACHING RESEARCH

Recently, Scruggs, Mastropieri, and McDuffie (2006) conducted a synthesis of qualitative research in the area of co-teaching. Using search procedures and selection criteria standards similar to those previously identified, 32 original reports of qualitative research in the area of co-teaching were identified, as identified in journal articles, ERIC documents, and dissertations. These reports described 323 teachers, including at least 138 special education teachers. These teachers were working in schools in the northeast, mid-Atlantic, southeast, midwest, west coast of the United States, and in Australia.

Besides geographical representation, identified co-teaching research focused on a range of grade levels: 13 reported on primary preschool or elementary school classrooms; 13 studied junior-high, middle-school, or high-school classrooms; while six reported on elementary and secondary classrooms combined. Schools and classrooms studied also represented a range of locations, including urban, suburban, and rural schools. Eleven of the reports identified “outstanding” examples of co-teaching, while others were described as more typical of the co-teaching experience. Thus, the samples identified may represent somewhat more successful collaborations than are found in schools in general.

After entry and data analysis, data were analyzed using procedures described previously. A total of 69 free nodes were identified, which were subsequently incorporated into four overriding categories:

1. Benefits of co-teaching.
2. Expressed needs for success in co-teaching.
3. Teacher roles in co-teaching.
4. How instruction is delivered in co-taught classes.

Benefits of co-teaching for students with disabilities, typically as represented by teacher and student reports and classroom observations, included individual attention, increasing self-esteem, and benefits of peer modeling. Students without disabilities were said to benefit from an enhanced understanding of others, and a developed sense of community and social responsibility. According to teachers, they benefited from enhanced learning about content knowledge or teaching/learning strategies, and felt general improvement in professional development.

Teachers also reported a number of needs for the success of co-teaching. These included administrative support, appropriate caseloads, co-teacher compatibility, time for planning, and appropriate student skill level for inclusive instruction.

A number of co-teaching roles was noted; however, the overwhelming number of co-teaching pairs could be characterized as ''one teach, one assist,'' where the general education teacher delivered the instruction and the special education teacher lent assistance to students throughout the class. Although some exceptions were noted, most researchers noted that the special education teacher assumed more of a secondary role. This appeared to be more true as the content became more specialized, although it was observed on both elementary and secondary levels.

Typically, the special education teacher interacted individually with students as needed, and modeled listening, note taking, and elaborations for students in the class. The special education teacher also provided assistance with class assignments, engaged in clerical tasks such as attendance, and applied discipline and behavior management strategies. However, these roles were typically confined to assistance with classroom tasks; providing intensive instruction in memory strategies and study skills, self-monitoring or self-management was only very rarely observed.

In summary, Scruggs et al. (2006) concluded that co-teaching has demonstrated some advantages to students and teachers. However, the instances observed in the classrooms are very different from those described in the literature, in that presently, special education teachers assume more of a "helper" role and the type of intensive strategy-based instruction often found in special education classrooms appeared to be lacking. Instead, a model where students are helped through their classroom experience appeared to prevail. Planning, compatibility of teachers, appropriate skill levels of students, and administrative support also appear to represent significant problems. Scruggs et al. suggested that such problems be considered carefully as schools move into the next phase of co-teaching. These conclusions appeared to carry some weight because they were based upon research involving a large number of teachers and classrooms throughout this country and Australia.

SUMMARY

The large accumulation of qualitative research in special education has provided a need for procedures for summarizing such research, in order to increase its external validity, and impact on the field of practice. Qualitative research synthesis methods have been implemented in other fields, largely health sciences, and could be applied with benefit to the field of special education.

One very promising technique for summarizing qualitative research is by treating each research report as an individual respondent, and coding and analyzing all such reports using qualitative data analysis methods and software, such as NVivo. Using this approach, a recent research synthesis revealed the state of research findings in co-teaching in inclusive classrooms, an area of study typically limited to isolated, small-scale investigations. Results from this study could be employed to make general recommendations for practice. It is hoped that with additional applications of meta-synthesis of qualitative research, qualitative research can begin to play a more prominent role in special education practice.

REFERENCES

Anzul, M., Evans, J. F., King, R., & Tellier-Robinson, D. (2001). Moving beyond a deficit perspective with qualitative research methods. *Exceptional Children*, *67*, 235–249.

Atkinson, P. (1995). Some perils of paradigms. *Qualitative Health Research*, *5*, 117–124.

Beck, C. T. (2001). Caring within nursing education: A metasynthesis. *Journal of Nursing Education*, *40*, 101–109.

Braden, J. P. (1992). Intellectual assessment of deaf and hard-of-hearing perple: A quantitative and qualitative research synthesis. *School Psychology Review*, *21*, 82–94.

Brantlinger, E., Jiminez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children*, *71*, 195–207.

Campbell, R., Pound, P., Pope, C., Britten, N., Pill, R., Morgan, M., & Donovan, J. (2003). Evaluating meta-ethnography: A synthesis of qualitative research on lay experiences of diabetes and diabetes care. *Social Science and Medicine*, *56*, 671–684.

Collet-Klingenberg, L. L. (1998). The reality of best practices in transition: A case study. *Exceptional Children*, *65*, 67–78.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, *52*, 291–302.

Downing, J. E., Ryndak, D. L., & Clark, D. (2000). Paraeducators in inclusive classrooms: Their own perceptions. *Remedial and Special Education*, *21*, 171–181.

Eisner, E. W. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.

Estabrooks, C. A., Field, P. A., & Morse, J. M. (1994). Aggregating qualitative findings: An approach to theory development. *Qualitative Health Research*, *4*, 503–511.

Ferguson, D. L. (1996). Is it inclusion yet? Bursting the bubbles. In: M. S. Berres, D. L. Ferguson, P. Knoblock & C. Woods (Eds), *Creating tomorrow's schools today: Stories of inclusion, change, and renewal* (pp. 16–37). New York: Teachers College Press.

Fraser, D. (1999). *QSR NUD*IST Vivo reference guide*. Melbourne, Australia: Qualitative Solutions and Research.

Freeman, D. (2000). Margaret Mead's *Coming of Age in Samoa* and Boasian culturalism. *Politics and the Life Sciences*, *19*, 101–103.

Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English-language learners. *Exceptional Children*, *66*, 454–470.

Glaser, B. G., & Strauss, A. L. (1971). *Status passage*. Chicago: Aldine.

Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research*, *50*, 438–460.

Jensen, L. A., & Allen, M. N. (1994). A synthesis of qualitative research on wellness–illness. *Qualitative Health Research*, *4*, 349–369.

Kozleski, E. B., & Jackson, L. (1993). Taylor's story: Full inclusion in her neighborhood elementary school. *Exceptionality*, *4*, 153–176.

LeCompte, M. D., & Preissle, J. (1993). *Ethnography and qualitative design in educational research*. New York: Academic Press.

Lincoln, Y. S. (1995). Emerging criteria for quality in qualitative and interpretive research. *Qualitative Inquiry*, *1*, 275–289.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Marks, S. U., Schrader, C., & Levine, M. (1999). Paraeducator experiences in inclusive settings: Helping, hovering, or holding their own? *Exceptional Children*, *65*, 315–328.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Noblit, G. W., & Hare, R. D. (1988). *Meta-ethnography: Synthesizing qualitative studies*. Newbury Park, CA: Sage.

Ogawa, B. T., & Malen, B. (1991). Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. *Review of Educational Research*, *61*, 265–286.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Salisbury, C. L., Palombaro, M. M., & Hollowood, T. M. (1993). On the nature and change of an inclusive elementary school. *Journal of the Association for Persons with Severe Handicaps*, *18*, 75–84.

Sandelowski, M., Docherty, S., & Emden, C. (1997). Qualitative meta-synthesis: Issues and techniques. *Research in Nursing and Health*, *20*, 365–371.

Schofield, J. W. (1990). Increasing the generalizability of qualitative research. In: E. W. Eisner & A. Peshkin (Eds), *Qualitative inquiry in education: The continuing debate* (pp. 201–232). New York: Teachers College Press.

Schwandt, T. A. (1996). Farewell to criteriology. *Qualitative Inquiry*, *2*, 58–72.

Scruggs, T. E., & Mastropieri, M. A. (1994). Successful mainstreaming in elementary science classes: A qualitative investigation of three reputational cases. *American Educational Research Journal*, *31*, 785–811.

Scruggs, T. E., & Mastropieri, M. A. (1995). Qualitative research methods in the study of learning and behavioral disabilities: An analysis of recent research. In: T. E. Scruggs & M. A. Mastropieri (Eds), *Advances in learning and behavioral disabilities* (Vol. 9, pp. 251–274). Oxford, UK: Elsevier.

Scruggs, T. E., & Mastropieri, M. A. (1996). Quantitative synthesis of survey research: Methodology and validation. In: T. E. Scruggs & M. A. Mastropieri (Eds), *Advances in learning and behavioral disabilities: Theoretical perspectives* (Vol. 10, Part A, pp. 209–223). Oxford, UK: Elsevier.

Scruggs, T. E., & Mastropieri, M. A. (2000). The effectiveness of mnemonic instruction for students with learning and behavior problems: An update and research synthesis. *Journal of Behavioral Education*, *10*, 163–173.

Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality*, *9*, 227–245.

Scruggs, T. E., Mastropieri, M. A., & McDuffie, K. (2006). *Co-teaching in inclusive classrooms: A synthesis of qualitative research*. Fairfax, VA: George Mason University, College of Education and Human Development.

Skinner, D., Bailey, D. B., Correa, V., & Rodriguez, P. (1999). Narrating self and disability: Latino mothers' construction of identities vis-à-vis their child with special needs. *Exceptional Children*, *65*, 481–495.

Weiss, M. P., & Lloyd, J. W. (2002). Congruence between roles and actions of secondary special educators in co-taught and special education settings. *Journal of Special Education*, *36*, 58–68.

West, J., & Oldfather, P. (1995). Pooled case comparison: An innovation for cross-case study. *Qualitative Inquiry*, *1*, 452–464.

Yin, R. K., & Heald, K. A. (1975). Using the case survey method to analyze policy studies. *Administrative Science Quarterly*, *20*, 371–381.

This page is left intentionally blank

SUBJECT INDEX

This page is left intentionally blank