School Psychology Review,
2005, Volume 34, No. 1, pp. 58-73

# Measurement Properties of Indirect Assessment Methods for Functional Behavioral Assessment: A Review of Research

Randy G. Floyd and Robin L. Phaneuf
*The University of Memphis*

Susan M. Wilczynski
*University of Nebraska Medical Center*

*Abstract.* Indirect assessment instruments used during functional behavioral assessment, such as rating scales, interviews, and self-report instruments, represent the least intrusive techniques for acquiring information about the function of problem behavior. This article provides criteria for examining the measurement properties of these instruments designed to identify functional relations and reviews 46 studies examining the use and interpretation of the Motivation Assessment Scale (Durand & Crimmins, 1992) or the Functional Assessment Interview (O'Neill et al., 1997). Results indicate at least three inadequacies in the research: (a) insufficient attention paid to evidence based on instrument content and informant response processes, (b) minimal evidence indicating consumer satisfaction and the singular or unique effects of instrument data on intervention development, and (c) excessive emphasis on research designs and measurement properties that assume stability in behavior–environment interactions across contexts.

Functional behavioral assessment (FBA) has received considerable attention in the professional literature of school psychology since the passage of the 1997 revision of the Individuals with Disabilities Education Act (IDEA, 1997), which mandated the use of FBA procedures and positive behavioral supports in some instances. Although IDEA did not specify what constitutes an FBA, a number of heuristic models for FBA have been proposed (e.g., O'Neill et al., 1997; Sugai, Lewis-Palmer, & Hagan-Burke, 2000; Witt, Daly, & Noell, 2000). Across these models, indirect methods of assessment are typically recommended during the beginning stages of an FBA. Indirect methods are characterized by being removed in time and place from the phenomena they measure (i.e., behaviors and functional relations). Examples include ratings of others, interviews, and self-report instruments (Cone, 1978).

Indirect methods of assessment offer a number of potential benefits to the process of FBA. For example, interviews may be useful in defining problem behaviors, determining their severity, and specifying optimal conditions under which to observe the behavior. Additionally, they may contribute to hypoth-

*Authors' Note.* We thank numerous instrument authors for responding to our requests for information, David M. Murray for assistance in locating resources describing categorical data-analytic procedures, and Renee Bergeron for proofing activities. We also appreciate editorial guidance from Cathy F. Telzrow and feedback from several anonymous reviewers about earlier versions of this article.

Correspondence concerning this article should be addressed to Randy G. Floyd, PhD, The University of Memphis, Department of Psychology, Memphis, TN 38152; E-mail: rgfloyd@memphis.edu.

Copyright 2005 by the National Association of School Psychologists, ISSN 0279-6015

esis-driven functional analysis[1] and reduce the time required to identify functional relations because fewer conditions may need to be manipulated. The informant-based nature of these methods also engages valuable stakeholders (e.g., parents and teachers) who will later be involved in the development, implementation, and monitoring of behavior support plans (Dunlap, Newton, Fox, Benito, & Vaughn, 2001). They logically are less time-consuming than direct observations, and they require notably less training, expertise, and staff collaboration to complete than functional analysis. Furthermore, they may be the only assessment method available when problem behaviors occur at a low frequency or when functional analysis is unethical or untenable (O'Neill et al., 1997).

Although the benefits of indirect methods for FBAs are numerous, informant reports stem from recollections of the problem behaviors and personal judgments about behavior–environment interactions and not their direct measurement. As a consequence, erroneous hypotheses may be developed or faulty conclusions may be drawn. Despite this significant limitation, no agreed-upon guidelines have emerged to evaluate the quality of the results obtained from indirect methods. Without a sound framework for examining their quality, consumers may be left not only misguided by their results but also uninformed about which instruments provide the most accurate and most useful information. Just as measurement standards are applied to the evaluation of norm-referenced tests to promote quality of results, so too must criteria be applied to indirect methods for FBAs.

Despite several publications focusing on evaluating measures used during FBAs (e.g., Cone, 1997, 1998; Gresham, 2003; Shriver, Anderson, & Proctor, 2001), there appears to be reluctance to embrace measurement standards for assessment of functional relations, particularly among individuals who embrace applied behavior analysis (Kratochwill & Shapiro, 2000). There are at least four reasons for this apparent reluctance. First, it can be argued that direct assessment of behavior through observation largely eliminates the need for evaluation of traditional measurement properties, such as construct or criterion-related validity (Ebel, 1961; Johnston & Pennypacker, 1993). Second, the ability of an FBA instrument to inform effective interventions (i.e., treatment utility) is viewed by some as the most important measurement property—if not ultimately the only important property (Hayes, Nelson, & Jarrett, 1987; Nelson-Gray, 2003). Third, the individual or idiographic nature of behavioral assessment does not appear to lend itself to group-level statistical analyses that are typically used to examine traditional measurement properties (Nelson, 1983; Nelson, Hay, & Hay, 1977). Finally, the methodology and research designs examining traditional measurement properties may be incongruent with what is known about behavior–environment interactions (Hartman, Roper, & Bradford, 1979; Nelson, 1983; Nelson et al., 1977). For example, because functional relations are expected to be variable across contexts—unlike traits targeted by many traditional measurement properties—they are less likely to demonstrate adequate levels of test–retest reliability or agree with other measures of functional relations obtained from different contexts.

These reasons for reluctance ground FBAs in low-inference assessment of directly observed behaviors and in assessments and interventions that are useful in addressing the target concerns for individuals. However, reluctance to develop and utilize measurement guidelines for assessment of functional relations hinders a broader understanding of which properties produce the best results and which properties produce the most error. Consideration of the measurement properties does not negate a practical focus on assessing and addressing the target concerns of individuals. Rather, measurement guidelines assist in evaluating the quality of assessment information in and of itself. Such guidelines ultimately lead users to develop stronger hypotheses and conclusions that are more accurate. Therefore, the purpose of this article is threefold. It offers research design characteristics and statistical analyses appropriate for examining assessment instruments designed to identify functional relations. It evaluates the research examining

the measurement properties of two such instruments. It also provides recommendations for future research and instrument use.

## Method

### Identification of Instruments

Several strategies were employed to identify indirect assessment instruments measuring functional relations to include in this review (Cooper, 1998; Lipsey & Wilson, 2001). First, electronic bibliographic searches of PsychInfo and ERIC (from January 1887 through August 2001) were conducted for published resources[2] describing indirect assessment instruments for FBA. Search terms included functional behavioral assessment, functional analysis, behavioral assessment, informant, questionnaire, behavior ratings scale, checklist, interview, psychometric properties, accuracy, validity, reliability, and generalizability. Second, after the resources were obtained, their content was reviewed to identify additional instruments measuring functional relations. Third, testing information clearinghouses on the World Wide Web (e.g., Buros Institute of Mental Measurements) and assessment instrument catalogs from prominent test publishers were searched. Finally, the primary authors of the instruments identified during the search were contacted by mail, electronic mail, or phone. They provided names of additional indirect assessment instruments to include in the review (Cooper, 1998). Based on these search strategies, 19 indirect assessment instruments (or collections of instruments) measuring functional relations were identified.

### Identification of Research

In order to identify all research supporting the use and interpretation of the identified instruments, additional electronic bibliographic searches of articles, books, and book chapters were conducted using PsychInfo and ERIC (from January 1887 to January 2003) using the names of the instruments and their authors' names. Subsequently, abstracts and reference lists obtained in the search were reviewed, and the texts of resources were scoured to identify primary research examining the use and interpretation of the instruments. In addition, during contact with instrument authors, research including the instruments and the existence of manuals or other published materials summarizing characteristics of the instruments were reported.

### Instrument Selection and Description

Instruments selected for review had at least (a) three studies examining their measurement properties, and (b) two studies conducted by independent researchers (i.e., those who were not authors of the instruments or associated with their research group). These two selection criteria were based on procedures designed to identify empirically supported therapies and interventions (Chambless & Hollon, 1998; Kratochwill & Stoiber, 2002; Wolery, 2003). The first criterion was used to include instruments with an established body of research examining their use. The second criterion was used to include research conducted across contexts, participants, and researchers. Only two instruments met both selection criteria.[3]

**Motivation Assessment Scale.** The MAS (Durand & Crimmins, 1992) is a rating scale designed to identify the function of problem behaviors. The MAS consists of 16 items scored on a 7-point scale (ranging from *Never* to *Always*) that may be completed independently or through an interview with parents, teachers, or careproviders. Items measure four classes of functional relations (Attention, Tangible, Escape, and Sensory), and the instrument yields four subscales representing these classes. Users identify functional relations based on the subscale with the highest rank.

**Functional Assessment Interview.** The FAI (O'Neill et al., 1997) is a lengthy interview for parents, teachers, or careproviders. The FAI is an extension of the Functional Analysis Interview (O'Neill, Horner, Albin, Storey, & Sprague, 1990), and it has recently been revised and abbreviated for use in school settings (Crone & Horner, 2003). The interview includes 11 sections soliciting information about problem behaviors, setting events, im-

mediate antecedents, consequences that may maintain the problem behaviors, and the efficiency of problem behaviors in obtaining reinforcement. Other sections facilitate identification of replacement behaviors, probable reinforcers, and discriminative stimuli. Additional sections include questions about the individual's communication skills and about previous interventions used to treat the problem behavior. The FAI concludes with the interviewer and interviewee constructing preliminary hypotheses regarding the functions of the problem behaviors and identifying alternate or replacement behaviors.

**Supporting research.** The search for research using the MAS or the FAI yielded 39 resources: 35 articles and four books. Three resources were excluded after full review because (a) the MAS was administered as a self-report form to children with mental retardation (Akande, 1994, 1998) and (b) only a general description of a case using the MAS was described (Lennox & Miltenberger, 1989). From the remaining resources, 46 studies were identified—35 examining the MAS and 11 the FAI. Although no studies were identified in the books, some validity evidence was presented in one book. Eight articles included more than one study, and two articles included studies that were coded for both the MAS and the FAI.

### Outline of Qualitative Analysis

**Development of review criteria.** Four sets of information were used to develop the review criteria for coding the research. The foundation of the criteria was the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Because the *Standards* provides only general guidelines for evaluation, three other sets of information were reviewed to identify guidelines that are more specific. These included (a) publications describing excellence in psychological measurement (e.g., Aiken, 1997; Anastasi & Urbina, 1997; Merrell, 2003; Messick, 1989, 1995; Nunnally & Bernstein, 1994) and (b) publications applying measurement standards to evaluation of behavior rating scales and intelligence and achievement tests (e.g., Bracken, 1987; Floyd & Bose, 2003; Hammill, Brown, & Bryant, 1992). A final type of source was recent publications describing the assumptions of assessment of behaviors and their functions, which were reviewed to tailor the criteria to these assumptions (e.g., Barnett, Lentz, & Macmann, 2000; Cone, 1977, 1997, 1998; Foster & Cone, 1995; Gresham, 1984, 2003; Gresham & Davis, 1988; Hayes et al., 1987; Haynes & O'Brien, 2000; Haynes, Richard, & Kubany, 1995; Johnston & Pennypacker, 1993; Shriver et al., 2001; Silva, 1993; Sturmey, 1994a).

Based on a synthesis of these four sets of information, coding criteria, coding sheets, and a codebook were constructed.[4] All materials were circulated among the authors and were piloted with five articles not included in this review. This process resulted in several revisions of the coding criteria and materials. The first section of the coding criteria focused on variables associated with the design and description of the studies, such as characteristics of participants being assessed, descriptions of the informants, and descriptions of the contexts in which the problem behaviors occurred.

The second section focused on the presence of research providing reliability and validity evidence and on the characteristics of the studies yielding this evidence. As apparent in Table 1, this section included (a) items for the presence of test–retest reliability analysis, the interval between ratings, and associated statistical analyses; (b) items for the presence of interrater reliability analysis and associated statistical analyses; and (c) items focusing on the five sources of validity evidence described in *Standards* (AERA, APA, & NCME, 1999). The presence of evidence based on test content, evidence based on response processes, and evidence based on consequences presented in Table 1 was coded. For evidence based on internal structure, the presence of statistical analyses was coded. For evidence based on relations with other variables, information was coded for (a) studies comparing the results of the MAS or FAI to other as-

sessment techniques or instruments and (b) studies where groups were compared using the MAS or FAI. For studies examining relations with other variables by comparing different measures (see Table 1), the other assessment techniques or instruments used in the comparison, informants, temporal proximity of the related measurements, and statistical analyses were coded. For the studies examining group differences, the manner in which groups were formed was coded. In addition to coding these variables associated with the measurement property, as evident in the right column in Table 1, several research design characteristics deemed necessary for valid conclusions or consistency with the assumptions of behavioral assessment were coded.

**Review process.** The first author reviewed and coded all 46 studies. To provide an index of interrater agreement, the second author reviewed and coded 30% of all studies (31% of the studies examining the MAS and 27% of the studies examining the FAI). Percentage agreement and the more conservative kappa statistic, which controls for agreement by chance, were calculated for items from both sections of the review criteria. For the section describing the design and description of the studies, there was 95% agreement, and kappa was .88. For the reliability and validity evidence section as a whole (in Table 1), there was 98.3% agreement, and kappa was .86. When items indicating presence of reliability and validity evidence were targeted and items nested under them (e.g., statistical analyses for test–retest reliability) were omitted, the percentage agreement for each of the seven general measurement properties presented in Table 1 was as follows: test–retest reliability (95%), interrater reliability (99.3%), evidence based on content (100%), evidence based on response processes (100%), evidence based on internal structure (97.7%), evidence based on relations with other variables (94.8%), and evidence based on consequences (100%). These estimates of interrater agreement indicate a high level of consistency in coding. All disagreements were evaluated and resolved by consensus.

## Results

### Participants Being Assessed

Across all studies, sample sizes ranged from 1 to 118 ($M = 20$, $Mdn = 5$). For the MAS, the mean sample size was 22.7 and the median 13. For the FAI, the mean sample size was 12.5 and the median 3. Participants ranged in age from children below 4 years of age (13 studies or 28.2%), children of preschool and school age (21 studies or 45.7%), adolescents (21 studies or 45.7%), to adults (19 studies or 41.3%). Diagnoses and classifications of participants were most frequently mental retardation (34 studies or 73.9%) and autistic disorder (21 studies or 45.7%). Two studies included children classified with emotional disturbance, 1 included a child with a learning disability, and 3 reported that children had no classification. Problem behaviors were most often self-injurious behaviors (34 studies or 73.9%), aggression (32 studies or 69.6%), and disruptive behaviors (27 studies or 58.7%). Across studies, 12 (26.1%) reported the actual frequencies of the problem behaviors, and 34 (73.9%) used either general descriptors to describe frequency (e.g., "frequent" or "problematic levels") or made no reference to the frequency of problem behaviors.

### Informants

Informants who most frequently responded to the indirect assessment instruments either independently or through an interview format were teachers or teacher assistants (19 studies or 41.3%), residential facility or developmental center staff (14 studies or 30.4%), and parents or guardians (10 studies or 21.7%). Across all studies, 16 (34.8%) quantified the extent of the informants' experience with the problem behaviors, and 30 (65.2%) used only general descriptors (e.g., "knowledgeable informants") or made no reference to the informants' experiences with the behaviors. Similarly, the vast majority of studies (36 or 78.3%) did not report the informants' training in behavioral assessment or behavioral principles. Only 4 studies (8.7%) reported that informants were trained in these areas, and only 6 studies

## Table 1
## Criteria for Coding Reliability and Validity Evidence

| Measurement Property | Variables Associated With Measurement Property | Research Design Characteristics |
|---|---|---|
| Test–retest reliability or stability | Presence of analysis<br>Interval<br>Short-term (< 1 month)<br>Long-term (≥ 1 month)<br>Statistical analyses<br>Correlations between *subscales* or groups of items<br>Correlations between *items*<br>Agreement between *subscales*, groups of items, or indicators of function<br>Agreement between *items* or individual question | Informant *did not* observe problem behavior between ratings |
| Interrater reliability | Presence of analysis<br>Statistical analyses (see test–retest reliability or stability above) | Experience of different raters controlled so that raters share all common experiences with problem behavior |
| Evidence based on content | Development of items based on literature review, review of case histories, or both<br>Development of items based on review of other instruments<br>Development of items based on direct observations<br>Development of items based on caregiver (parent, teacher, staff) input<br>Development of items based on expert opinion (other than authors)<br>Review and evaluation of item clarity<br>Review and evaluation of bias (gender, racial, special needs)<br>Review and item tryouts by users in applied settings<br>Readability analysis of items<br>Readability analysis of written informant instructions<br>Questions about function grouped by experts or caregivers | |
| Evidence based on response processes | Evaluation of instrument instructions and response formats<br>Interviews with informants about thought processes<br>"Think aloud" protocols with informants<br>Analysis of match between interviewer verbalizations and administration criteria | |

(Table 1 continues)

(Table 1 continued)

| Measurement Property | Variables Associated With Measurement Property | Research Design Characteristics |
|---|---|---|
| Evidence based on internal structure | Item intercorrelations within scales or subgroups of items<br>Item intercorrelations across all items (within total score)<br>Item correlations with total score (i.e., item–total correlations)<br>Internal consistency coefficients including only items within scales or subgroups of items<br>Internal consistency coefficients including all items (within total score)<br>Exploratory factor analysis of items/Confirmatory factor analysis of items<br>Correlations between subscales or subgroups of items | Informant blind to the results of other assessment methods |
| Evidence based on relations with other variables | Results of the instrument were compared to results from a distinctly different assessment instrument or technique that identifies function<br>Presence of analysis<br>Instrument or technique (and informant, if applicable)<br>Timing of measurements<br>Concurrent<br>Predictive<br>Statistical analyses<br>Correlations between subscales, groups of items, or indicators of function<br>Agreement between subscales, groups of items, or indicators of function<br><br>Groups were formed based on an external characteristic and results of the instrument were compared between groups<br>Presence of analysis<br>Grouping<br><br>Grouped by classification of participant on whom data are collected<br>Grouped by problem behavior or response class | |
| Evidence based on consequences | Evaluation of treatment utility<br>Evaluation of acceptability of instrument or assessment method/technique for purpose of assessment<br>Evaluation of use of instrument in relation to intended outcome<br>Cost–benefit analysis<br>Cost–effectiveness analysis | Instrument used alone to assess function<br>Instrument used with a larger assessment battery to assess function<br>Not reported if instrument used alone or in larger battery |

(13.0%) reported the general level of training or education of informants.

### Context of Problem Behaviors

Informants most often assessed problem behaviors occurring in school or preschool settings (21 studies or 45.7%) and in residential treatment or clinical settings (17 studies or 37.0%). Although it is likely that informants routinely delineated the specific contexts or settings (e.g., classroom) in which the problem behaviors occurred before completing the FBA, only 16 studies (34.8%) reported this information. Even fewer studies (8 or 17.4%) reported the specific situations in which the problem behaviors occurred, such as during independent reading time or during small group instruction.

### Measurement Properties

**Reliability.** Five studies (10.9%) examined test–retest reliability or stability. All of these studies focused on the MAS; none focused on the reliability of the functions identified using the FAI. All but one of these studies used samples of children, and all but two were conducted in school settings. Of these five studies, four examined short-term generalizability across time (<1 month), and three examined long-term stability (≥1 month). Statistics used to examine test–retest reliability or stability included correlations between *subscales* or groups of items (three studies); correlations between *items* (three studies); formulas examining agreement between subscales, groups of items, or indicators of function (one study); and formulas examining agreement between items or individual questions (three studies). Only two studies reported the specific settings in which the problem behaviors occurred, and only one reported specific situations. None of the studies reported controlling the informants' exposure to the problem behaviors between ratings. Thus, even in the studies that reported the specific contexts of the problem behaviors, error attributed primarily to the passage of time between ratings cannot be identified.

Eighteen studies (39.1%) examined interrater reliability. Of the 18 studies, 15 examined the administration or scoring of the MAS, and 3 examined the FAI. Statistics used to examine interrater reliability included correlations between subscales or groups of items (13 studies) and correlations between items (8 studies). Also evident were studies examining agreement between subscales, groups of items, or indicators of function (11 studies) and those examining agreement between items or individual questions (6 studies). None of the studies reported controlling the observations of informants so that their experiences with the problem behaviors were identical. However, 4 studies reported the specific settings in which the problem behaviors occurred, and 1 reported specific situations. Thus, for most all studies reviewed, it is likely that informants observed problem behaviors in different settings and situations and, as a result, completed the assessment based on observations of different behavior–environment interactions. Despite the sizeable number of studies examining this measurement property, error attributed primarily to different perceptions and memories of informants cannot be readily identified.

**Evidence based on content.** Across the resources for the MAS and FAI, little information was located documenting the appropriateness and completeness of instrument items or questions. As described very briefly in only two resources, the content of the MAS items is based on a strong research base examining the environmental contingencies maintaining the problem behaviors of children and adults with developmental disabilities. Items were developed and piloted over a 4-year period, and items were added, deleted, or altered during item tryouts with the aid of careproviders of children with mental retardation, autistic disorder, and other developmental disabilities. However, the procedure for assigning items to subscales is unclear. No formal item-sorting procedures drawing upon the decisions of experts appear to have been utilized for assigning items to subscales. There were no descriptions of consultation with external content experts during the development and evaluation of items, no evaluation of item bias, and no evaluation of item or instruction readability or clarity.

It is apparent that the FAI has gone through several stages of development (see O'Neill et al., 1990; O'Neill et al., 1997) and that it is thorough in leading informants to describe and prioritize problem behaviors, to identify variables affecting behavior, and to promote the development of hypotheses. Thus, it appears to be based on expert opinion and feedback from users. However, this review revealed no description of its development or evidence based on content to support its validity.

**Evidence based on response processes.** When the characteristics of the informant and the interaction between the informant and the assessment situation were considered, no evidence was found. This finding is similar to other behavior rating scales and behavioral assessment interviews (Beaver & Busse, 2000; Floyd & Bose, 2003).

**Evidence based on internal structure.** Internal relations between items from the MAS have been investigated through item-level factor analysis and internal consistency analysis. More specifically, four analyses examined the factor structure of the MAS using exploratory factor analysis, and five analyses examined the relations between items within each of its four scales. Most studies included only adolescents and adults with mental retardation displaying self-injurious behavior and other problem behaviors. No studies used confirmatory factor analysis. Although meaning is most frequently derived from the MAS via identifying its highest ranked scale, three analyses examined the statistical relations among *all items*, which is more appropriate for an instrument measuring the severity of a wide variety of problem behaviors (e.g., Achenbach, 1991).

**Evidence based on relations with other variables.** The review yielded 36 analyses (across 25 studies) comparing the results of the MAS or FAI to other instruments or techniques yielding information about functional relations. From the studies comparing the MAS to other indirect assessment instruments, 2 analyses compared it to other rating scales (i.e., the Problem Behavior Questionnaire [Lewis, Scott, & Sugai, 1994] and the Questions About Behavioral Functioning Scale [Vollmer & Matson, 1999]), 3 analyses compared it to semistructured interviews (e.g., the FAI), and 2 analyses compared it to direct questioning about functional relations. For the FAI, 3 analyses compared it to rating scales, 1 analysis compared it to another semistructured interview, and 1 analysis compared it to direct questioning about functional relations. Notably, 14 analyses (10 for the MAS, 4 for the FAI) compared the functional relations identified by the MAS or FAI to those identified through functional analysis. Ten analyses compared their results to functions identified through direct observation. Of these analyses including direct observation, 9 used data collected using an antecedent–behavior–consequence (ABC) format. These comparative analyses most frequently examined the agreement between subscales, groups of items, or other indicators of function (30 or 83.3%). Only 6 analyses (16.7%) used correlations between measures. Although the number of analyses examining this measurement property is impressive, very few studies containing them (9 or 19.6%) reported that informants were blind to the results of the other assessment instruments and techniques. In addition, only 15 analyses (41.6%) reported the specific settings in which the problem behaviors occurred, and only 5 (13.8%) reported specific situations.

**Evidence based on consequences.** Evidence of the treatment utility of the MAS and the FAI was identified in nine studies (19.6%).[5] However, it is challenging to determine the unique contribution of the instruments to intervention development from these studies because, in all cases but one, the instruments were used as part of larger assessment batteries measuring functional relations. No studies addressing acceptability of these instruments and their outcomes have been published to date. In addition, no resource was identified that included a cost–benefit or cost–effectiveness analysis.

## Discussion

A number of problems plague the existing research examining the measurement properties of FBA measures. In addition to the ab-

sence of well-designed research examining some properties, the research designs and statistical analyses used in many studies appear to be incongruent with the assumptions of the assessment of behavior and its function. In addition, other research design characteristics and statistical analyses appear to have been flawed.

## Implications for Research and Evaluation

Well-designed research is needed to provide sound evidence supporting the use and interpretation of indirect assessment instruments for FBAs. Such research will be improved if authors and journal editors include details in publications pertaining to (a) the assumptions of behavioral assessment, such as variability of behavior according to setting and situation, and (b) research designs and statistical analyses that promote greater precision in identifying the measurement characteristics of these instruments. As evident from the sizeable number of studies included in this review that included fewer than 10 participants, researchers need not abandon a focus on individuals in the search for general principles related to measurement properties of instruments.

**Content and response processes.** This review indicates that researchers have invested too little effort identifying and reporting evidence based on test content and response processes. Although the content of most indirect assessment instruments for FBA is drawn from the rich research base of applied behavioral analysis, it is important that authors make explicit the resources used and steps taken during instrument development. Potential consumers and content experts independent of the development could assist in refining instrument wording and ensuring that information about all relevant variables, such as setting events, is elicited (Haynes et al., 1995).

A number of naturalistic and analogue studies could examine the response processes of informants completing these instruments. For example, some informant characteristics (e.g., training in observation of behavior and its function) could be systematically manipulated to examine their effects on the accuracy of resulting data. Informants also could be asked to verbalize their internal speech (i.e., "think aloud") while completing instruments, and verbal protocols could be analyzed to identify components deemed necessary for accurate identification of functional relations and apparent failure to use these components (Ericsson & Simon, 1993). Researchers examining FBA interviews may benefit from use or modification of existing coding schemes for verbalizations during interviews, such as the Consultant Analysis Record (Bergan & Kratochwill, 1990; Bergan & Tombari, 1971). For example, research could examine verbal interactions and their correspondence to interview content to determine integrity of assessment procedures.

**Internal structure.** Although statistical analysis of the internal relations between segments of most interviews for FBAs, such as the FAI, do not appear to be tenable, the application of such analysis to items from rating scales and select self-report instruments seems appropriate—if evidence based on test content supports item groupings. This review revealed five analyses examining internal structure of the MAS. It is logical to expect that some groups of items (i.e., subscales) would covary in a meaningful manner. However, there would be no reason to expect internal relations across all items (e.g., for a total score) because groups of items representing each function are designed to measure distinct (and largely mutually exclusive) environment–behavior interactions. Increasing the number of well-constructed items within a related group and conducting item-level analysis with large groups of informants early in the development of the instrument would likely be beneficial. An unpublished revision of the MAS seems to have benefited from both more items and initial item-level analysis (Durand, 1997; Haim, 2003).

**Relations with other variables.** More carefully designed studies should be conducted to examine the correspondence between measures of functional relations. To increase the integrity of results, research designs should

ensure that assessment data are collected in a manner in which informants, observers, or other experimenters are blind to the results of other assessment methods. This review indicated that fewer than 20% of the studies providing evidence based on external relations reported this effort. Continued attention should also be paid to statistical analyses appropriate for categorical-level data. In addition to raw agreement indexes, the use of other statistical analyses for determining agreement between categorical-level data, such as kappa coefficients, tests of marginal homogeneity, and generalizability coefficients, also can be explored (Kraemer, Periyakoil, & Noda, 2002; Li & Lautenschlager, 1997; Powers & Xie, 2000). As evident from this review, Pearson product-moment correlations between measures may continue to be used when data, such as items, are continuous. Spearman correlations may be used when the probability of functions indicated by instruments can be rank ordered.

Research examining external relations of FBA instruments should continue to anticipate the confounding influences of context on measurement of functional relations. Most assessment instruments focusing on symptoms of psychopathology (e.g., Achenbach, 1991; Reich, 2000) yield the most meaningful results when informants observe symptoms across contexts. However, the truest measures of functional relations result from deliberate specification and observation of settings (e.g., a classroom) and situations (e.g., independent seat work) in which the behaviors occur (Shriver et al., 2001). This review identified few studies that reported information about the specific contexts of the problem behaviors.

**Consequences of assessment.** Although treatment utility is considered by some to be the ultimate measurement property, this review revealed only one study that examined the isolated effects of the MAS. In all other cases, the instruments were included in a larger assessment battery to inform the development of interventions. Research should partial out the contributions of specific assessment results to treatment development (Hayes et al., 1987; Nelson-Gray, 2003). This research can be accomplished through carefully controlled analogue studies or through meticulous documentation of intervention development based on a stepwise review of assessment results.

This review also revealed that research examining the MAS and FAI has not examined the degree to which consumers are satisfied with the assessment process and their outcomes (Eckert & Hintze, 2000; Gresham & Lopez, 1996; Wolf, 1978). If consumers do not find an assessment instrument acceptable, they may be unlikely to actively engage in the assessment process or to trust and utilize the assessment results. In addition, the costs and benefits of use of these instruments have not been quantified and compared (Yates & Taub, 2003). For example, the use of staff time is a variable that is likely to drive the choice of assessment instruments—especially under conditions where there is limited information about the quality of the instruments (Shriver & Proctor, 2004). The recent evolution of the FAI into an abbreviated interview is evidence that instrument authors are attentive to this variable (Crone & Horner, 2003).

**Consistency across time and informant.** Because behaviors may serve different functions in different contexts, research focusing on test–retest reliability and interrater reliability should examine the consistency of reports of functional relations derived from observations of identical settings and situations. As a result of the variability of behavior and its function across contexts and time, informants' repeated observations of behaviors in applied settings between initial and subsequent completion of the instruments undermines accurate test–retest reliability estimates. Similarly, informants observing different contexts in which the behaviors occur weaken estimates of inter-rater reliability. Influences on consistency of assessment results across time and informant will be best pinpointed by carefully controlled laboratory or analogue studies, such as those including observation of problem behaviors by videotape (Johnston & Pennypacker, 1993). In addition, study of these measurement properties will benefit from greater attention paid to statistical analyses appropriate for categorical-level data.

## Limitations

This review is not without limitations. Although efforts were made to ensure inclusion of all related studies, it is possible that the search strategies for the instruments and the research examining them missed important resources. Similarly, inclusion of instruments that did not meet the criteria for inclusion, such as some self-report inventories, may have led to different conclusions. Finally, although inter-rater agreement in coding the studies was high, it is possible that the decision-making was consistent but biased in a critical manner.

## Implications for Practice

Given the limitations of indirect assessment instruments used during FBAs and the restricted and sometimes flawed research examining the two most often studied instruments, school psychologists should continue to rely on direct, minimally inferential assessment techniques to identify functional relations. However, indirect assessment instruments, such as comprehensive interviews, conducted with knowledgeable informants may yield unique and relevant information that aids in the identification of functional relations. For example, setting events or person variables (e.g., establishing operations) affecting problem behaviors may not be apparent if observations, functional analysis, or both techniques are used exclusively. If indirect assessment instruments are used, school psychologists should carefully review the content of the instruments due to minimal evidence based on content identified in this review. They should determine if content is applicable to the settings of interest, if important information is elicited, and if items are worded well and appropriately for informants.

Practitioners can and should contribute to the research base regarding these assessment instruments. Although some measurement properties are most applicable to carefully controlled research and group-level analyses, accumulated study of individual cases of FBA can also be useful in building understanding of these instruments, especially about response processes, treatment utility, and user satisfaction. In particular, to demonstrate evidence of response processes, practitioners can study adherence to interviews to determine whether they are being completed as intended and whether specific variables, such as training scripts, increase assessment integrity (Ehrhardt, Barnett, Lentz, Stollar, & Reifin, 1996). Practitioners can describe the effects of well-monitored interventions developed solely from indirect assessment instruments and examine the incremental benefits of more direct assessment on interventions. Practitioners can also assess the acceptability of assessment instruments through surveys of teachers, parents, and other consumers of these instruments.

## Footnotes

[1] In this article the term *functional analysis* is used to refer to the process of systematically altering antecedents of behavior, consequences of behavior, or both to examine effects on problem behavior.

[2] Throughout this article, the term *study* is used to describe a systematic investigation examining data from a set of participants. The term *resource* is used to describe the medium for these studies and for other evidence of measurement properties. Examples include journal articles, books, and book chapters. The term *analysis* is used to describe the part of a study in which statistical techniques are used to answer a research question. For example, one resource, an article, contained four studies. One of these studies contained two analyses examining relations with other variables.

[3] At least a pair of studies examining the measurement properties of the Questions About Behavioral Functioning Scale (Vollmer & Matson, 1999), the Functional Assessment Informant Record for Teachers (Edwards, 2002), the Problem Behavior Questionnaire (Lewis, Scott, & Sugai, 1994), the Student-Assisted Functional Assessment Interview (Kern, Dunlap, Clarke, & Childs, 1994), and the Student-Directed Functional Assessment Interview (O'Neill et al., 1997) were identified. However, none were supported by sufficient independent research.

[4] The coding sheets and coding manual can be obtained from the first author.

[5] Treatment utility evidence was defined liberally as substantiation of links between assessment results and resulting interventions (cf. Hayes et al., 1987 and Nelson-Gray, 2003).

## References

References marked with an asterisk were included in the literature review.


Achenbach, T. M. (1991). *Integrative guide for the 1991 CBCL/4–18, YSR, and TRF profiles.* Burlington: University of Vermont, Department of Psychiatry.

Aiken, L. R. (1997). *Psychological testing and assessment* (9th ed.). Needham Heights, MA: Allyn & Bacon.

Akande, A. (1994). The motivation assessment profiles of low-functioning children. *Early Child Development and Care, 101,* 101–107.

Akande, A. (1998). Some South African evidence of the inter-rater reliability of the Motivation Assessment Scale. *Educational Psychology, 18,* 111–115.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice–Hall.

*Arndorfer, R. E., Miltenberger, R. G., Woster, S. H., Rortvedt, A. K., & Gaffaney, T. (1994). Home-based descriptive and experimental analysis of problem behaviors in children. *Topics in Early Childhood Special Education, 14,* 64–87.

Barnett, D. W., Lentz, F. E., Jr., & Macmann, G. (2000). Psychometric qualities of professional practice. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 355–386). New York: Guilford Press.

* Barton-Arwood, S. M., Wehby, J. H., Gunter, P. L., & Lane, K. L. (2003). Functional behavior assessment rating scales: Intrarater reliability with students with emotional or behavioral disorders. *Behavioral Disorders, 28,* 386–400.

Beaver, B. R., & Busse, R. T. (2000). Informant reports: Conceptual and research bases of interviews with parents and teachers. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 257–287). New York: Guilford Press.

Bergan, J. R., & Kratochwill, T. R. (1990). *Behavioral consultation and therapy.* New York: Plenum Press.

Bergan, J. R., & Tombari, M. L. (1971). The analysis of verbal interactions occurring during consultation. *Journal of School Psychology, 13,* 209–225.

* Bihm, E. M., Kienlen, T. L., Ness, M. E., & Poindexter, A. R. (1991). Factor structure of the Motivation Assessment Scale for persons with mental retardation. *Psychological Reports, 68,* 1235–1238.

* Bird, F., Dores, P., Moniz, D., & Robinson, F. (1989). Reducing severe aggression and self-injurious behavior with functional communication training: Direct, collateral and generalized results. *American Journal of Mental Retardation, 94,* 37–48.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5,* 313–326.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8,* 427–430.

Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy, 9,* 882–888.

Cone, J. D. (1997). Issues in functional analysis in behavioral assessment. *Behavioral Research and Therapy, 35,* 259–275.

Cone, J. D. (1998). Psychometric considerations: Concepts, contents, and methods. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed., pp. 22–46). Boston: Allyn & Bacon.

* Conroy, M. A., Fox, J. J., Bucklin, A., & Good, W. (1996). An analysis of the reliability and stability of the Motivation Assessment Scale in assessing the challenging behaviors of persons with developmental disabilities. *Education and Training in Mental Retardation and Developmental Disabilities, 31,* 243–250.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3rd. ed). Thousand Oaks, CA: Sage.

* Crawford, J., Brockel, B., Schauss, S., & Miltenberger, R. G. (1992). A comparison of methods for the functional assessment of stereotypic behavior. *Journal of the Association for Persons with Severe Handicaps, 17,* 77–86.

* Crone, D. A., & Horner, R. H. (2003). *Building positive behavior support systems in schools: Functional behavioral assessment.* New York: Guilford Press.

* Cunningham, E., & O'Neill, R. E. (2000). Comparison of results of functional assessment and analysis methods with young children with autism. *Education and Training in Mental Retardation and Developmental Disabilities, 35,* 406–414.

* Duker, P. C., & Sigafoos, J. (1998). The Motivation Assessment Scale: Reliability and construct validity across three topographies of behavior. *Research in Developmental Disabilities, 19,* 131–141.

Dunlap, G., Newton, J., S., Fox, L., Benito, N., & Vaughn, B. (2001). Family involvement in functional assessment and positive behavior support. *Focus on Autism and Other Developmental Disabilities, 16,* 215–221.

* Dunlap, G., White, R., Vera, A., Wilson, D., & Panacek, L. (1996). The effects of multi-component, assessment-based curricular modifications on the classroom behavior of children with emotional and behavioral disorders. *Journal of Behavioral Education, 4,* 481–500.

Durand, V. M. (1997). Motivation Assessment Scale II: Test version. Unpublished instrument.

* Durand, V. M., & Carr, E. G. (1991). Functional communication training to reduce challenging behavior: Maintenance and application in new settings. *Journal of Applied Behavior Analysis, 24,* 251–264.

* Durand, V. M., & Crimmins, D. B. (1988). Identifying the variables maintaining self-injurious behavior. *Journal of Autism and Developmental Disorders, 18,* 99–117.

* Durand, V. M., & Crimmins, D. B. (1992). *The Motivation Assessment Scale administrative guide.* Topeka, KS: Monaco & Associates.

* Durand, V. M., Crimmins, D. B., Caulfield, M., & Taylor, J. (1989). Reinforcer assessment: I. Using problem behavior to select reinforcers. *Journal of the Association for Persons With Severe Handicaps, 14,* 113–126.

* Durand, V. M., & Kishi, G. (1987). Reducing severe behavior problems among persons with dual sensory impairments: An evaluation of a technical assistance model. *Journal of the Association for Persons with Severe Handicaps, 12,* 2–10.

Ebel, R. (1961). Must all tests be valid? *American Psychologist, 15,* 546–553.

Eckert, T. L., & Hintze, J. M. (2000). Behavioral conceptions and applications of acceptability: Issues related to service delivery and research methodology. *School Psychology Quarterly, 15,* 123–148.

Edwards, R. P. (2002). A tutorial for using the Functional Assessment Informant Record for Teachers. *Proven Practice: Prevention and Remediation Solutions for Schools, 4,* 31–33.

Ehrhardt, K. E., Barnett, D. W., Lentz, F. E., Jr., Stollar, S. A., & Reifin, L. H. (1996). Innovative methodology in ecological consultation: Use of scripts to promote treatment acceptability and integrity. *School Psychology Quarterly, 11,* 149–168.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.

Floyd, R. G., & Bose, J. E. (2003). A critical review of rating scales assessing emotional disturbance. *Journal of Psychoeducational Assessment, 21,* 43–78.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248–260.

* Gibson, A. K., Hetrick, W. P., Taylor, D. V., Sandman, C. A., & Touchette, P. (1995). Relating the efficacy of naltrexone in treating self-injurious behavior to the Motivation Assessment Scale. *Journal of Developmental and Physical Disabilities, 7,* 215–220.

Gresham, F. M. (1984). Behavioral interviews in school psychology: Issues in psychometric adequacy and research. *School Psychology Review, 13,* 17–25.

Gresham, F. M. (2003). Establishing the technical adequacy of functional behavioral assessment: Conceptual and measurement challenges. *Behavioral Disorders, 28,* 282–298.

Gresham, F. M., & Davis, C. J. (1988). Behavioral interviews with teachers and parents. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 455–493). New York: Guilford Press.

Gresham, F. M., & Lopez, M. F. (1996). Social validation: A unifying concept for school-based consultation research and practice. *School Psychology Quarterly, 11,* 204–227.

Haim, A. (2003). The analysis and validation of the Motivation Assessment Scale-II Test Version: A structural equation model. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 63*(10-B): 4903.

Hammill, D. D., Brown, L., & Bryant, B. R. (1992). *A consumer's guide to tests in print* (2nd ed.). Austin, TX: PRO–ED.

* Haring, T. G., & Kennedy, C. H. (1990). Contextual control of problem behavior in students with severe disabilities. *Journal of Applied Behavior Analysis, 23,* 235–243.

Hartman, D. P., Roper, B. L., & Bradford, C. C. (1979). Some relationships between behavioral and traditional assessment. *Journal of Behavioral Assessment, 1,* 3–21.

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42,* 963–974.

Haynes, S. N., & O'Brien, W. H. (2000). *Principals and practice of behavioral assessment.* New York: Kluwer Academic/Plenum.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

*Individuals with Disabilities Education Act Amendments of 1997,* (Pub L No. 105–17. 20 USC Chapter 33, Sections 1400 *et seq.* (Statute)

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

* Kearney, C. A. (1994). Interrater reliability of the Motivation Assessment Scale: Another, closer look. *Journal of the Association for Persons with Severe Handicaps, 19,* 139–142.

Kern, L., Dunlap, G., Clarke, S., & Childs, K. E. (1994). Student-assisted functional assessment interview. *Diagnostique, 19*(2,3), 29–39.

* Kinch, C., Lewis-Palmer, T., Hagan-Burke, S., & Sugai, G. (2001). A comparison of teacher and student functional behavior assessment interview information from low-risk and high-risk classrooms. *Education and Treatment of Children, 24,* 480-494.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine, 21,* 2109–2129.

Kratochwill, T. R., & Shapiro, E. S. (2000). Conceptual foundations of behavioral assessment in schools. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 3–15). New York: Guilford Press.

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly, 17,* 341–389.

* Lawry, J. R., Storey, K., & Danko, C. D. (1993). Analyzing problem behaviors in the classroom: A case study of functional analysis. *Intervention in School and Clinic, 29,* 96-100.

Lennox, D. B., & Miltenberger, R. G. (1989). Conducting a functional assessment of problem behavior in applied settings. *Journal of the Association of Persons with Severe Handicaps, 14,* 304–311.

Lewis, T. J., Scott, T., & Sugai, G. (1994). The Problem Behavior Questionnaire: A teacher-based assessment instrument to develop interventions to develop functional hypotheses of problem behavior in general education classrooms. *Diagnostique, 19,* 103–115.

* Lewis, T. J., & Sugai, G. (1996). Functional assessment of problem behaviors: A pilot investigation of the comparative and interactive effects of teacher and peer social attention on students in general education settings. *School Psychology Quarterly, 11,* 1–19.

Li, M. F., & Lautenschlager, G. (1997). Generalizability theory applied to categorical data. *Educational and Psychological Measurement, 57,* 813–822.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Merrell, K. W. (2003). *Behavioral, social, and emotional assessment of children and adolescents.* Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validity of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Nelson, R. O. (1983). Behavioral assessment: Past, present, and future. *Behavioral Assessment, 5,* 195–206.

Nelson, R. O., Hay, I. R., & Hay, W. M. (1977). Comments on Cone's "The relevance of reliability and validity for behavioral assessment." *Behavior Therapy, 8,* 427–430.

Nelson-Gray, R. O. (2003). Treatment utility of psychological assessment. *Psychological Assessment, 15,* 521–531.

* Newton, J. T., & Sturmey, P. (1991). The Motivation Assessment Scale: Interrater reliability and internal consistency in a British sample. *Journal of Mental Deficiency Research, 35,* 472–474.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

* O'Neill, R. E., Horner, R. H., Ablin, R. W., Storey, K., & Sprague, J. R. (1990). *Functional analysis of problem behaviors: A practical guide.* Sycamore, IL: Sycamore.

* O'Neill, R. E., Horner, R. H., Ablin, R. W., Sprague, J. R., Storey, K., & Newton, J. S. (1997). *Functional assessment and program development for problem behaviors: A practical handbook.* New York: Brooks/Cole.

* Paclawskyj, T. R., Matson, J. L., Rush, K. S., Smalls, Y., & Vollmer, T. R. (2001). Assessment of the convergent validity of the Questions About Behavioral Function scale with analogue functional analysis and the Motivation Assessment Scale. *Journal of Intellectual Disability Research, 45,* 484–494.

Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis.* San Diego, CA: Academic Press.

* Reese, R. M., Richman, D. M., Zarcone, J., & Zarcone, T. (2003). Individualizing functional assessments for children with autism: The contribution of perseverative behavior and sensory disturbances to disruptive behavior. *Focus on Autism and Other Developmental Disabilities, 18,* 89–94.

Reich, W. (2000). Diagnostic Interview for Children and Adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 59–66.

* Scotti, J. R., Schulman, D. E., & Hojnacki, R. M. (1994). Functional analysis and unsuccessful treatment of Tourette's syndrome in a man with profound mental retardation. *Behavior Therapy, 25,* 721–738.

* Shogren, K. A., & Rojahn, J. (2003). Convergent reliability and validity of the Questions About Behavioral Function and the Motivation Assessment Scale: A replication study. *Journal of Developmental and Physical Disabilities, 15,* 367–375.

Shriver, M. D., Anderson, C. M., & Proctor, B. (2001). Evaluating the validity of functional behavior assessment. *School Psychology Review, 30,* 180–192.

Shriver, M. D., & Proctor, B. (2005). *Social validity in disseminating functional behavior assessment technology: Research directions for meeting consumer needs.* Manuscript submitted for publication.

* Sigafoos, J., Kerr, M., & Roberts, D. (1994). Interrater reliability of the Motivation Assessment Scale: Failure to replicate with aggressive behavior. *Research in Developmental Disabilities, 15,* 333–342.

Silva, F. (1993). *Psychometric foundations and behavioral assessment.* Newbury Park, CA: Sage.

* Singh, N. N., Donatelli, L. S., Best, A., Williams, E. E., Barrera, F. J., Lenz, M. W., Landrum, T. J., Ellis, C. R., & Moe, T. L. (1993). Factor structure of the Motivation Assessment Scale. *Journal of Intellectual Disabilities Research, 37,* 65–75.

* Spreat, C., & Connelly, L. (1996). Reliability analysis of The Motivation Assessment Scale. *American Journal of Mental Retardation, 100,* 528–532.

Sturmey, P. (1994a). Assessing the function of aberrant behaviors: A review of psychometric instruments. *Journal of Autism and Developmental Disorders, 24,* 293–304.

* Sturmey, P. (1994b). Assessing the functions of self-injurious behavior: A case of assessment failure. *Journal of Behavior Therapy and Experimental Psychiatry, 25,* 331–336.

Sugai, G., Lewis-Palmer, T., & Hagan-Burke, S. (2000). Overview of the functional behavioral assessment process. *Exceptionality, 8*(3), 149–160.

* Thompson, S., & Emerson, E. (1995). Inter-informant agreement on the Motivation Assessment Scale: Another failure to replicate. *Mental Handicap Research, 8,* 203–208.

Vollmer, T. R., & Matson, J. L. (1999). *Questions About Behavioral Function manual.* Baton Rouge, LA: Scientific Publishers.

Witt, J. C., Daly, E. M., & Noell, G. (2000). *Functional assessments: A step-by-step guide to solving academic and behavior problems.* Longmont, CO: Sopris West.

Wolery, M. (2003, October). *Single-subject methods: Utility for establishing evidence for practice.* Paper presented at the 19th Annual International Conference on Young Children with Special Needs and Their Families, Washington, DC.

Wolf, M. M. (1978). The case of subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11,* 203–214.

* Yarbrough, S. C., & Carr, E. G. (2000). Some relationships between informant assessment and functional analysis of problem behavior. *American Journal of Mental Retardation, 105,* 130–151.

Yates, B. T., & Taub, J. (2003). Assessing the costs, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, here's how. *Psychological assessment, 15,* 478–495.

* Zarcone, J. R., Rodgers, T. A., Iwata, B. A., Rourke, D. A., & Dorsey, M. F. (1991). Reliability analysis of the Motivation Assessment Scale: A failure to replicate. *Research in Developmental Disabilities, 12,* 349–360.

Randy G. Floyd, PhD is an Assistant Professor of Psychology at The University of Memphis. He received his doctoral degree in School Psychology from Indiana State University. His research interests include assessment of cognitive abilities, identification of reading and mathematics aptitudes, and improving behavioral assessment methods.

Robin L. Phaneuf, PhD is an Assistant Professor of Psychology at The University of Memphis. She received her doctoral degree in School Psychology from the University of Massachusetts Amherst. Her research interests include assessment and intervention for early learning and social emotional behavior of young children and application of school-based consultation procedures to community settings.

Susan M. Wilczynski, PhD, BCBA, is an Assistant Professor of Pediatric Psychology at Munroe-Meyer Institute for Genetics and Rehabilitation, a unit of the University of Nebraska Medical Center. She is a licensed psychologist and certified behavior analyst whose applied and research interests address treatment of children with Autistic Spectrum Disorders. She developed and administrates Project BEST-CASE, an intensive early childhood intervention program for children with autism and related disorders.