Evaluating the Validity of Functional Behavior Assessment

Mark D. Shriver
University of Nebraska Medical Center

Cynthia M. Anderson
West Virginia University

Briley Proctor
Florida State University

Abstract. This article provides a brief overview of functional behavior assessment and contrasts and compares functional behavior assessment and traditional psychological assessment. Messick's (1995) theory of construct validity is presented and then applied to the evaluation of the validity of functional behavior assessment measures. Finally, implications and guidelines of current construct validity theory are presented for school psychologists developing and/or using FBA measures and procedures.

Following the most recent passage of the Individuals with Disabilities Education Act (IDEA, 1997), functional behavior assessment has received increasing attention in the educational and school psychology literature (Donovan, 1999; Dwyer, 1998; Nelson, Roberts, Bullis, Albers, & Ohland, 1999; Steege, 1999; Steege & Northup, 1998; Stewart, 1999). Although relatively new in education, functional behavior assessment has always been an integral component of applied behavior analysis (Baer, Wolf, & Risley, 1968; Neef, 1994; Skinner, 1953). As educators and school psychologists continue to define and delineate of what functional behavior assessment will consist for their school districts, it will be necessary for them to look to the applied behavior analysis literature for empirically supported methods and technology in functional behavior assessment.

Functional behavior assessment (FBA) incorporates many of the same assessment methods as more traditional psychological assessment. Interviews, behavior rating forms, and observations are often key components of an FBA (e.g., Hintze & Eckert, 2001; O'Neil et al., 1997). However, the conceptual underpinnings of FBA differ substantially from those of more traditional psychological assessment (Hayes, Nelson, & Jarrett, 1986), and for this reason, applied behavior analysts have not typically attended to many of the psychometric issues common to traditional psychological measurement (e.g., internal reliability, content validity, construct validity). These psychometric principles of measurement are well known to many school psychologists and are commonly used to evaluate traditional methods of assessment.

School psychologists typically receive minimal training in applied behavior analysis (Shriver & Watson, in press). Therefore, it may be difficult for school psychologists to evalu-
ate the different FBA procedures and measures that are currently available given that these procedures were not developed within the same conceptual framework to which many psychometric principles of measurement are applicable. To provide guidance on evaluating FBA, this article provides a brief overview of FBA, contrasts and compares FBA and traditional psychological assessment, presents a model of validity developed for the evaluation of traditional psychological assessment data, and applies the validity model to the evaluation of FBA data. Finally, the implications for school psychologists developing and evaluating FBA measures and methods for their school district are presented.

**Functional Behavior Assessment**

FBA refers to the process of gathering information to determine relations between variables, particularly functional relations. The term *functional* refers to a demonstration of the effect of one variable on another variable, in other words cause-effect relations (Skinner, 1938, 1953). A description of a functional relation may be correlational, although the term *functional* in the behavior analytic literature is usually reserved for cause-effect relations that have been experimentally demonstrated (Carr, 1994; Cone, 1997). Methods used to develop hypotheses regarding functional relations include interviews with significant others (e.g., parents, teachers), interviews with the client (i.e., student), behavior rating forms completed by others or the client, direct observation, and systematic manipulation of environmental events or contingencies (Iwata, Dorsey, Slifer, Bauman, & Richman, 1982/1994; O’Neill et al., 1997; Steege & Northup, 1998). The term *functional analysis* is often used to refer to procedures that systematically manipulate environmental events to determine a functional relation (Cone, 1997; O’Neill et al., 1997).

Behavioral assessment is an assessment paradigm that differs conceptually from traditional psychological assessment (Hayes, Nelson, & Jarrett, 1986), and FBA may be considered to be under the umbrella of behavioral assessment (Bellack & Hersen, 1988; Cone, 1978). However, assessment practices within the paradigm of behavioral assessment typically do not identify cause-effect relations. In fact, it is common for behavioral assessment methods to have behavior data (i.e., frequency, intensity, duration) as the primary product, rather than functional (i.e., cause-effect) data. FBA refers specifically to the identification of cause-effect relations (Cone, 1997; Hayes, Nelson, & Jarrett, 1986). Review of the history of inclusion of FBA in IDEA suggests that this definition of FBA (i.e., identification of cause-effect relations) is what was intended by the law.

To identify functional relations between the behavior of an individual and the individual’s environment, it is necessary to include the individual’s target behavior and environmental events and the relation between the behavior and environmental event in the FBA process. Assessment methods and procedures that only identify or assess the target behavior (or components thereof) or environmental events without assessment of the relation between environmental events and target behavior are not considered functional behavior assessments, although the information may be useful in an FBA.

Identifying cause-effect relations in the interaction of the individual and his or her environment is a stated goal of FBA because, ultimately, knowledge of the cause-effect relation is used to design interventions leading to behavior change. Prediction and control of behavior are important outcomes of FBA (Baer, Wolf, & Risley, 1968; Skinner, 1953, 1974). FBA data that do not lead to prediction and control of behavior are not perceived to be meaningful data (i.e., data that have social and treatment validity).

FBA in applied behavior analysis has tended to focus largely on the reinforcing consequences of behavior (Carr, 1994). However, antecedent conditions (e.g., discriminative stimuli, establishing operations, setting events) are also important (DeGrandpre, 2000; Dunlap, Kern-Dunlap, Clarke, & Robbins, 1991; Gresham, Watson, & Skinner, this issue; McGill, 1999; Smith & Iwata, 1997; Van Camp et al., 2000), and are receiving increased attention as important components in a comprehensive FBA (e.g., Horner, 1994; O’Neill et al., 1997). In addition, it should be noted that
contrary to frequently stated perceptions (e.g., Miller, 2000), behavior analysis does encompass important emotional, cognitive, and phenomenological variables (Skinner, 1953). Radical behaviorism (as a philosophical approach to the science of human behavior), and behavior analysis (as the science of human behavior) provide empirically based theoretical and methodological guidelines for the study of complex human behavior and learning (e.g., DeGrandpre, 2000; Donahoe & Palmer, 1994; Sidman, 2000; Skinner, 1974).

Traditional Psychological Assessment

In contrast to FBA, traditional psychological (and psychoeducational) assessment often involves a process of gathering information to determine patterns of behavior that are inferred to represent a particular construct or syndrome within the child. Constructs are typically used in psychology because they allow for classification or grouping of similar instances of behavior and allow people to efficiently communicate in concise terms what has been witnessed (Crocker & Algina, 1986). Constructs frequently become more than a label for a group of behaviors, however, as they often come to be viewed as explanations for the occurrence of the behaviors. According to Anastasi (1982), “Each construct is developed to explain and organize observed response consistencies” (p. 144). Thus, the construct that was once simply a label for a group of behaviors is now viewed as the cause of the behaviors.

Relations between environmental events and child behavior may be described in a traditional psychological assessment; however, the inferred construct being measured (i.e., aggression, attention, learning style) is hypothesized to mediate or cause the behavior-environment relationship. These constructs or mediating variables, which are viewed as causal agents (independent variables) in traditional psychological assessment, are not evaluated as such in an FBA. Rather, the FBA focuses on the behaviors that originally contributed to the development of the construct. For example, the construct of “aggression” might include such behaviors as hitting, swearing, or threatening gestures, as well as behaviors that are more difficult to measure such as homicidal thoughts or increased heart rate. These behaviors would be targeted in an FBA. In other words, in an FBA the overt and covert behaviors that make up structures are events to be explained (dependent variables), not part of the explanation (Carr, 1993; Moore, 1994; Skinner, 1974).

The types of psychoeducational assessment commonly employed by school psychologists (Reschly & Wilson, 1995) often involve examining scores from norm-referenced measures to determine how similar or different an individual is relative to a normative sample. Causes for behavior are inferred from an analysis of score patterns within and across assessment measures (Flanagan & McGrew, 1997; Kaufman, 1994). In essence, traditional psychoeducational assessment asks the question, “Does this child look (i.e., score) similar to or different from other children who exhibit behaviors that make up a particular construct (e.g., learning disability, Attention Deficit Hyperactivity Disorder)”? Subsequently, the construct is inferred to cause the presenting problem (e.g., poor grades, noncompliance). Therefore, the outcome of a traditional psychoeducational evaluation typically consists of data hypothesized to represent a construct inferred to have a particular effect on an individual’s behavior or inferred to have a relationship with a separate construct (e.g., aggression is related to impulsivity).

The Construct Validity of Assessment Data

The above-mentioned differences between FBA and traditional psychological assessment focused on the data produced by the two approaches because the products, not the methods, of assessment are of primary interest (Goodwin & Goodwin, 1999). Therefore, the differences in the type of data obtained from the two assessment approaches will affect how the data are evaluated (Cone, 1998). In FBA, the data represent a cause-effect relation, whereas in traditional psychological assessment the primary data are a score or pattern of scores that are hypothesized to represent a construct within the child.

Traditional psychological assessment is concerned with the measurement of hypotheti-
ical constructs. Therefore, demonstrating evidence for the reliability and validity of the assessment data is important so that one may have confidence that the results of the assessment have some degree of consistency or reproducibility and accurately represent the construct assumed to exist within the child. The evaluation of traditional psychological assessment, therefore, has focused on how much error the scores contain (i.e., reliability) and how well the scores represent the construct of interest (i.e., validity; Goodwin & Goodwin, 1999). Providing evidence for the construct validity of the data is the goal of traditional psychometric evaluation.

The Functional Relation Construct

It may be said that the “construct” under consideration in an FBA is the functional relation. Hypotheses regarding a functional relation are developed based on data from indirect measures (i.e., no manipulation of variables) such as interviews, rating forms, and naturalistic observations. Similarly, hypotheses regarding functional relations in the natural environment also occur based on data from more direct FBA procedures (i.e., systematic manipulation of variables) such as analog functional analyses. There are two fundamental differences, however, between the construct of functional relation and most other psychological constructs. First, the functional relation exists in the interaction between the behavior of the child and the child’s environment, not within the child. Second, a functional relation can be directly demonstrated to exist, whereas psychological constructs cannot.

Unless directly demonstrated, there is a level of inference involved in hypotheses regarding functional relations based on FBA data. The level of inference inherent in a hypothesis based on FBA data is directly affected by the degree of similarity between the environment in which an FBA is conducted and the natural environment in which the functional relation occurs. Although it can reasonably be asserted that the level of inference in an FBA is substantially lower than in traditional psychological assessment (Martens, 1992), it is still imperative to determine the accuracy of FBA data. Accuracy reflects the “extent to which observed values approximate the ‘true’ state of nature” (Johnston & Pennypacker, 1993, p. 138). Accuracy is similar to the notion of representational validity (Cone, 1997), also referred to as trait validity (Campbell & Fiske, 1959, in Cone, 1997). Establishing evidence for the accuracy of FBA data also establishes evidence for its construct validity.

Although the products of traditional psychological assessment and FBA differ, the goal of evaluation of the assessment data for both assessment paradigms is to establish the construct validity of the data. Current psychometric theory for establishing evidence for construct validity may assist in the evaluation of FBA methods.

Messick’s Theory of Construct Validity

It has been argued that the traditional trinity of validity (content, criterion-related, construct) is artificial and essentially meaningless (Goodwin & Goodwin, 1999), and it has been proposed that all forms of reliability and validity evidence contribute to the construct validity of a measure’s outcome. For example, Messick (1995) proposes that all of the currently used psychometric concepts can be subsumed as being part of six aspects of construct validity. As a unified construct, “[construct] validation is an empirical evaluation of the meaning and consequences of measurement” (Messick, 1995, p. 747). Although developed within a traditional psychological assessment paradigm, Messick’s theory appears to have application for the evaluation of the validity of FBA data. Below, six aspects of construct validity proposed by Messick (1995) are presented and briefly reviewed. Note that these six areas are not meant to be separate processes for gathering evidence for construct validity, but are meant to reflect differing aspects of a unified construct validation process. The application of each of the areas to the evaluation of FBA is then presented and examples are provided.

Content Relevance and Representativeness

The evaluation of content relevance and representativeness is most closely related to the traditional concept of content validity. The pr-
primary concern is in demonstrating that the assessment (a) adequately samples a broad range of items representing the construct and/or (b) contains a broad enough sample of behavior to establish that the assessment is not too narrowly focused. In addition, it is important to demonstrate that the items on the measure or the samples of behavior represent the most meaningful or important aspects of the construct. Finally, the technical quality of the items must also be evaluated.

There are three considerations in evaluating the content relevance and representativeness of an FBA. First, the content and format of the assessment procedure should assist in selecting and operationalizing a target behavior as well as assisting in the identification of a competing or replacement behavior. It is important to identify target behaviors and events that are observable and measurable. In addition, FBA data should lead to effective treatment. This typically means that problems are not only resolved, but competencies are developed. So if an FBA reveals that a child's physical aggression (e.g., hitting teacher) is reinforced by temporary escape from academic tasks (e.g., the child is sent to the principal's office after hitting others), an intervention based on FBA data should reduce the physical aggression and teach a more positive skill (e.g., asking for assistance or a break) so the child may manage difficult tasks more appropriately. In addition to, or as an alternative to, teaching a new skill, FBA data may suggest an intervention involving environmental manipulations such as providing the child with extra tutorial assistance or changing the level of difficulty of the academic tasks to better match the child's skill level.

Second, it is important that developers of FBA measures and procedures provide evidence that the assessment adequately samples a broad range of environmental events that may lead to a functional relation, and/or the measure contains a broad enough sample of potential causative events that it can be established that the assessment is not too narrowly focused. For example, for the child who is aggressive and is escaping tasks or demands, it is important that the FBA assess not only the consequences when demands are presented (e.g., escape, attention), but also the level and type of demand that precedes the target behavior and absence of the target behavior (e.g., many versus few directions, math assignments versus social studies assignments, easy tasks versus difficult tasks).

Third, it is important to demonstrate that the samples of behavior and events represent aspects of the potential functional relations that are relevant for the setting for which it is to be used. For example, the content and format of an FBA interview for a teacher may be very different from the content and format of an FBA interview for a direct service provider in a residential facility for adults with severe disabilities (see Table 1 for a summary of standards in evaluating the validity of an FBA).

Substantive Theories, Process Models, and Process Engagement

The evaluation of substantive theories, process models, and process engagement is also similar to content validity, but goes further in suggesting that not only should the content of the measure appear to cover or adequately sample the construct domain, it should also reflect those components of the construct to which an individual or group of individuals are most likely to respond. This implies that there are relevant and irrelevant components of a construct that may be measured, and the assessor needs empirical evidence that the samples of behavior or items on a particular measure are the most relevant or most important aspects of a construct. For example, if the target behavior is a habit, such as thumb-sucking in the classroom, then it is likely that the function of the behavior is what has been termed automatic reinforcement. If the target behavior is disruptive behavior in the classroom (e.g., shouting, out-of-seat), then it is likely that the function of the behavior is attention (from peers and/or teacher) or escape from demands. Similarly, classroom lighting, noise, number of students in the class, or the child's sleep schedule may all be functionally related to a child's disruptive behavior in the class, but not necessarily the most relevant variables (e.g., type of demand, peer attention) or the aspects of the functional relation that can be most effectively and efficiently controlled. The primary point is
Table 1
Standards for Evaluating the Validity of a Functional Behavior Assessment

<table>
<thead>
<tr>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Does the assessment assist with selecting and operationalizing a target behavior and a competing</td>
</tr>
<tr>
<td>or replacement behavior?</td>
</tr>
<tr>
<td>2. Does the assessment sample a broad range of environmental events that may lead to a functional</td>
</tr>
<tr>
<td>relation? Can it be established that the assessment is not too narrowly focused?</td>
</tr>
<tr>
<td>3. Are the samples of environmental events relevant for the setting in which the assessment is</td>
</tr>
<tr>
<td>taking place?</td>
</tr>
<tr>
<td>4. Are the behaviors targeted and the environmental events sampled variables that the individual is</td>
</tr>
<tr>
<td>most likely to encounter in the setting in which the assessment is taking place and variables that</td>
</tr>
<tr>
<td>are most probable in a functional relation?</td>
</tr>
<tr>
<td>5. Does the assessment identify a cause-effect relation that can lead to effective treatment?</td>
</tr>
<tr>
<td>Process</td>
</tr>
<tr>
<td>6. Are there empirical data demonstrating the existence and treatment validity of the functional</td>
</tr>
<tr>
<td>relations purported to be assessed by the measure or procedure?</td>
</tr>
<tr>
<td>Reliability and Generalizability</td>
</tr>
<tr>
<td>7. Is there documented evidence for the internal, test-retest, and interrater reliability of the</td>
</tr>
<tr>
<td>measure/procedure appropriate for the context and time frame in which the assessment is expected to</td>
</tr>
<tr>
<td>be used?</td>
</tr>
<tr>
<td>8. Is there documented evidence for the generalizability of the assessment outcome across scorers,</td>
</tr>
<tr>
<td>behavior, time, setting, method of assessment, and dimension appropriate to the context and time</td>
</tr>
<tr>
<td>frame in which the assessment is to be used?</td>
</tr>
<tr>
<td>Convergent Correspondence with External Variables</td>
</tr>
<tr>
<td>9. Is there documented evidence for the accuracy of the measure? For example, has the outcome of</td>
</tr>
<tr>
<td>the measure been compared with the outcome of a single-subject experimental design (functional</td>
</tr>
<tr>
<td>behavior analysis) or has the outcome of the measure been compared with the outcome of another</td>
</tr>
<tr>
<td>measure that has demonstrated accuracy?</td>
</tr>
<tr>
<td>Consequences</td>
</tr>
<tr>
<td>10. Is there documented evidence for the treatment validity of the assessment data?</td>
</tr>
<tr>
<td>11. Is there documented evidence for the social validity of the assessment data?</td>
</tr>
</tbody>
</table>

that the assessment measure or procedure should assist in narrowing potential relevant functional variables that the individual is most likely to encounter specific to the targeted behavior in the setting in which the assessment is taking place. Most important, the end result of the FBA process should be the identification of a cause-effect relation that can lead to effective treatment.

**Scoring Models as Reflective of Task and Domain Structure**

In traditional psychological assessment, the scoring models for a particular measure should correspond to the theoretical or hypothesized structure of the construct. For example, subscale scores should be consistent with the underlying structure of the construct, which may be determined through empirical means (e.g., factor analysis) or through theory.

The outcome of an FBA will be a functional relation, not a score. Scoring models are not particularly applicable in this case. Conceptually, however, the functional relation should reflect current empirically based theory on cause-effect relations. If new functions or cause-effect relations are proposed to be mea-
sured (i.e., identification of constructs inferred to affect behavior), then empirical data directly demonstrating the existence and utility of the cause-effect relation should be presented. For example, the effect of positive and negative reinforcement on behavior has been clearly established through numerous empirical studies in the experimental and applied behavior analysis research literature. In addition, the effects of antecedent variables on behavior, such as stimulus control and establishing operations, have been empirically demonstrated in previous research (McGill, 1999; Skinner, 1953; Smith & Iwata, 1997). As an applied example, if in a particular teacher’s classroom a child’s talking out of turn is frequently followed by teacher attention, which is functionally reinforcing, then simply being in that classroom or in sight of that teacher might come to evoke the behavior of talking out of turn. Other events that are more distal to the behavior may also play an important role. For example, teacher attention may not be as reinforcing for a child who is lacking sleep relative to days the child has received enough sleep.

**Generalizability and the Boundaries of Score Meaning**

The representativeness of the content of a measure affects the degree to which one can have confidence that its results are generalizable to other samples of the same construct. Likewise, knowledge or hypotheses of a construct’s stability across time or between raters should correspond to the measure’s results over time, situations, and across raters. This aspect of construct validity appears to encompass the principles of reliability and generalizability of assessment data.

In classical test theory, it is assumed that because of its indirect nature, all measurement contains error. That is, the construct of interest (e.g., aggression) is measured indirectly through observations or ratings of behavior (e.g., behavior rating scales). Reliability refers to the consistency or reproducibility of the results obtained from an assessment method. Methods of establishing reliability are means of determining the amount of measurement error present under different conditions. Test scores with high reliability have low amounts of error and vice versa. Reliability, in turn, is a necessary (but not sufficient) condition for validity. In classical test theory, types of reliability include internal consistency, equivalent (or alternate) forms, test-retest, interrater, and the standard error of measurement (SEM). The degree of reliability present in a set of scores generated from a test provides confidence that similar results would be found (i.e., generalize) over different occasions, different raters, or different samples of the same performance domain (Linn & Gronlund, 1995).

The traditional methods of studying reliability described above provide an overall estimate of the amount of error contained in test scores, but they are not capable of partitioning the error into its various components. Generalizability theory is a more powerful means of studying reliability because it provides a differentiated error component (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In other words, Generalizability theory can be used to examine the degree to which the outcomes of an assessment generalize across “universes” including scorers, behavior, time, setting, method of assessment, and dimension (Cone, 1977). Generalizability theory has been noted to have utility in the evaluation of both traditional psychological assessment and behavioral assessment outcomes (Cone, 1977; Crocker & Algina, 1986; Gresham & Lambros, 1998). A recent applied example of the utility of Generalizability theory is supplied by Hintze and colleagues who used this approach in evaluating the quality of outcomes from curriculum-based assessment, a type of behavioral assessment (Hintze, Owen, Shapiro, & Daly, 2000).

The empirical basis for Generalizability theory is based on group data (Crocker & Algina, 1986; Gresham & Lambros, 1998; Hintze et al., 2000), but conceptually the theory is consistent with the idea in behavioral assessment that there should be relative levels of agreement across these universes of generalizability (Cone, 1977; Gresham & Lambros, 1998). It should be noted, however, that discussion of Generalizability theory in the behavioral assessment literature has focused
primarily on the generalizability of behavior data and scores (e.g., Cone, 1977; Hintze et al., 2000) not on the generalizability of environmental and/ or functional relation data, which is necessary in the evaluation of FBA data.

Because of the situational and idiosyncratic nature of functional relation data, the reliability or generalizability of the data may be difficult to demonstrate. It may be expected, however, that an FBA that is repeated in the same setting and within a relatively short time frame should produce consistent findings (i.e., test-retest reliability). Also, if a different (but reliable) measure was obtained in the same setting and at the same time as the original measure, then both measures should identify similar relations (alternate form reliability). Finally, if two observers independently complete the same FBA measure in the same setting and within a relatively short time frame (preferably at the same time), they should identify similar functional relations (interobserver reliability).

It is necessary for developers of FBA procedures or measures to provide evidence for the types of reliability mentioned above. In addition, developers should demonstrate how the items or components of the measure relate to each other and contribute to the outcome (e.g., internal consistency). To date, there is little evidence regarding the reliability of some of the more widely known FBA measures (Sturmey, 1994). Behavior analysis and assessment has typically focused more on demonstrating the “believability” of data using interobserver agreement (Johnston & Pynpach, 1993). However, interrater agreement is not equivalent to reliability (Goodwin & Goodwin, 1999).

Similar to reliability, generalizability is an important consideration in the evaluation of an FBA outcome. Also similar to reliability, expecting high levels of correlation across all the dimensions (e.g., raters, time, settings) may be unrealistic given the nature of the functional relation. It is important to have some level of stability across universes of generalization, however, if there is to be any hope that assessment results will translate into positive treatment results. Some researchers are beginning to examine and demonstrate the generalizability of different forms of FBA (i.e., analog functional analysis, structured interviews) across time, raters, and settings (Anderson, Freeman, & Scotti, 1999; Horner, Day, & Day, 1997; Lerman & Iwata, 1993; Mace, Lalli, & Lalli, 1991).

Convergent and Discriminant Correlations with External Variables

Evaluating the convergent and discriminant correlations of a measure with external variables is an external aspect of construct validity and refers to how the data of the target measure correlate with the data of other measures that are purported to measure the same, similar, or different constructs. The theory supporting the construct suggests how the scores obtained on one measure should relate to those on the other measure, and relationships found in the expected direction are evidence of validity.

The most salient point of this aspect of construct validity for FBA is providing evidence that the functional relation identified by the FBA is the same functional relation identified by another FBA measure with demonstrated accuracy. The evaluation of the accuracy of data from FBA requires two elements: (a) an incontrovertible index of behavior and (b) rules and procedures for using the measure (Cone, 1998). Establishing rules and procedures for a measure is self-explanatory, but finding an incontrovertible index of behavior is much more difficult to determine and attain. Cone (1997, 1998) suggests that the outcome of a measure be compared with the outcome of another measure that has been determined empirically and/or through consensus to represent the target behavior. When evaluating the accuracy of FBA outcomes, the systematic, experimental manipulation of environmental events (i.e., functional analysis) is considered the most accurate measure of a functional relation. That is, other methods of functional behavior assessment (interviews, rating forms, naturalistic observations) often are compared to a functional analysis when evaluating their accuracy (e.g., Anderson et al., 1999; Lerman & Iwata, 1993). A functional analysis is currently the best “proof” of the accuracy of hypotheses regarding functional relations.
The systematic, experimental manipulation of environmental events (i.e., functional analyses) has received much attention in applied behavior analysis, particularly for evaluating and treating severe behavior problems (e.g., self-injurious behavior, aggression, tantrums) exhibited by individuals with developmental disabilities. Although much of the research literature has used analog settings to conduct functional analyses, functional analysis may also take place in natural settings (Derby et al., 1992; Northup et al., 1991), with common behavior problems (Ervin, DuPaul, Kern, & Friman, 1998; Northup et al., 1995; Repp & Karsh, 1994) and academic behavior problems (Cooper et al., 1992).

Functional analyses rely on experimental single subject design methodology. The assumptions underlying single subject experimental designs differ significantly from the assumptions underlying group based statistical designs (Barlow & Hersen, 1984). An understanding of group based inferential statistical design is necessary in evaluating the reliability and validity of traditional psychological measures. Likewise, it is necessary to understand single subject experimental designs to evaluate the validity of FBA. There are many good texts (e.g., Barlow & Hersen, 1984), chapters (e.g., Anderson & Kim, in press), and review articles (e.g., Polaha & Allen, 1999) on single subject design. Two issues of importance for evaluating the quality of single subject designs are internal and external validity.

Internal validity refers to the power of an experimental design to isolate the independent variable that is affecting the dependent variable (Barlow & Hersen, 1984). This gets to the heart of demonstrating a functional relation and directly affects the accuracy of the assessment. Behavior is multiply determined, meaning that there are usually many variables that directly affect a child's behavior and learning. These variables include setting events, establishing operations, discriminative stimuli, and reinforcing and punishing consequences. The effect of any one variable on behavior is relative to all the other variables affecting behavior (Martens & Kelly, 1993; Shriver & Kramer, 1997). Group based statistical designs cannot identify functional relations; therefore, the utility of single subject experimental designs in the accurate identification of multiple functions and/or in the assessment of the relative contribution of multiple variables continues to be researched and evaluated (e.g., Call, Casey, Reed, Wacker, & Barretto, 2000).

External validity refers to whether the results of an experimental design generalize across domains such as clients, settings, therapists/teachers, and time (Barlow & Hersen, 1984). A functional analysis is externally valid to the extent that the results obtained accurately describe the functional relation that exists in the target or natural environment. The more similar the variables are between the setting in which the functional analysis takes place and the natural environment, the more likely an accurate functional relation will be identified. For example, a functional relation identified through analog functional analysis in a clinical setting has a greater probability of being inaccurate for a student's classroom behavior compared to the results of a functional analysis conducted in that student's classroom. Additional research on the generalizability of analog functional analysis is needed, as well as additional research on the application of functional analysis procedures in natural settings. In addition, the fact that behavior is often multiply determined means that although we may identify a functional relation in a functional analysis, that functional relation may not be the one that leads to effective treatment in the natural environment (Anderson et al., 1999). It is suggested that multiple forms of FBA be conducted to identify all important functional relations (Carr, Yarbrough, & Langdon, 1997).

**Consequences as Validity Evidence**

The sixth aspect of validity represents the most global conceptualization of validity in that it moves beyond internal and external evidential pieces to examine how the real and potential consequences of assessment results are both interpreted and used. Measurement results are used to make many different types of decisions about individuals and groups, and the intended and unintended consequences of
using results may have both positive and negative effects on an individual, group, or system. The construct and the construct's measure should have logical and empirical support for the decisions made based on the results of a measure (Messick, 1995).

The evaluation of the consequences of FBA is the aspect of construct validity that corresponds most closely with the principles of treatment validity and social validity in behavior analysis. Demonstrating that an FBA addresses concerns that are socially important and that the FBA results contribute to effective treatment planning is essential and has been a primary focus of evaluating FBA from the behavior analysis perspective.

Treatment validity refers to whether assessment data contribute to effective intervention development (Hayes, Nelson, & Jarrett, 1987). Treatment validity has long been a primary consideration in applied behavior analysis (Baer et al., 1968). An FBA method may have demonstrated reliability and accuracy, but if the data do not contribute to developing effective treatment then the assessment method is not considered useful. Hayes, Nelson, and Jarrett (1986, 1987) provide useful guidelines for evaluating the treatment utility of assessment data. In addition, the wealth of published interventions based on the results of an FBA (see almost any issue of the Journal of Applied Behavior Analysis) helps speak to the treatment validity of FBA.

It is not only important that assessment data contribute to effective intervention planning. The data must also address concerns that are socially valid or socially meaningful (Baer, Wolf, & Risley, 1968). Determining the social validity of FBA is largely determined by the people conducting the assessment and by societal consensus, although multiple perspectives on the assessment of social validity exist (cf. Geller, 1991). Social validity also means that the efficiency of the FBA relative to its effectiveness needs to be evaluated. It is easier to conduct an FBA interview than to conduct direct observations or systematic manipulations of potentially important environmental variables in the natural environment without specific training. However, data from the functional analysis or from direct observations may be more likely to lead to more effective treatment relative to the data from an interview. As with all assessment methods, the costs and benefits of FBA continue to require examination. Finally, it is important that developers of FBA measures and procedures demonstrate how to use the data from the FBA to develop treatment.

Implications for School Psychologists

Although there are similarities between traditional psychological assessment and FBA in some of the assessment methods (i.e., interviews, behavior rating forms, direct observations), there are substantial differences in the other assessment procedures (standardized norm-referenced testing versus single subject experimental design). The two assessment processes were developed within different assessment paradigms and produce fundamentally different types of data. Subsequently, evaluation of the assessment data has historically differed in emphasis (reliability and validity versus treatment validity and social validity). Messick's (1995) model of construct validity assists in bridging some of the conceptual gaps between the evaluation of the two assessment approaches and may be useful when evaluating the validity of FBA.

It is important that school psychologists keep in mind that functional behavior assessments assist in the identification of cause-effect relationships that lead to treatment development. An assessment process that leads to data that do not identify a cause-effect relation is not a functional behavior assessment. It is also imperative that those responsible for evaluating functional behavior assessments be knowledgeable about the strengths and limitations of single case experimental designs as that is the base upon which FBA data will ultimately be evaluated.

Table 1 summarizes standards to consider when developing and evaluating FBA measures and procedures. Note that the essential framework for construct validity that is proposed by Messick (1995) is retained for organization of the standards. We are not aware of any FBA measure currently available that meets all the standards presented. Future re-
search in the evaluation of FBA should look toward addressing the questions and standards presented in Table 1.

The fact that additional research is needed on establishing the validity of FBA, particularly in classroom settings, does not mean that school psychologists should refrain from using FBA. Rather, it is recommended that school psychologists become knowledgeable and skilled in FBA procedures that are developed within the paradigm of applied behavior analysis. As noted in the introduction, FBA is a core component of behavior analysis and has a long and successful history within that philosophical and scientific paradigm. It is expected, therefore, that FBA measures and procedures developed, utilized, and evaluated in applied behavior analysis will be most likely to meet many of the standards presented for determining the validity of the FBA data. FBA measures developed within behavior analysis will be most likely to meet validity standards related to (a) the identification of cause-effect relations (Item 5 on Table 1), (b) a basis in empirical literature (Item 6), (c) treatment validity (Item 10), and (d) social validity (Item 11).

Applied behavior analysis has provided a strong empirical base for the evaluation of the validity of FBA, and future research incorporating current theory in construct validity will assist school psychologists in effectively using empirically valid functional behavior assessment.

References


Call, N., Casey, S., Reed, G. K., Wacker, D. P., & Barretto, A. (2000, May). Correlation between a concurrent op-
erants assessment and functional analysis. Poster presented at the 2000 convention of the Association for Behavior Analysis, Washington, DC.


Individuals with Disabilities Education Act Amendments of 1997 (PL 105-17). 20 USC Chapter 33 (Sections 1400 et seq).


Mark Shriver is an Associate Professor in Pediatrics and Psychology at the Munroe-Meyer Institute for Genetics and Rehabilitation and the University of Nebraska Medical Center. He is co-editor of the journal *Proven Practice: Prevention and Remediation Solutions for Schools.* His clinical, research, and training interests include parent training, pediatric school psychology, academic and behavioral assessment, and applied behavior analysis in the schools.

Cynthia Anderson is an Assistant Professor of Psychology in the Behavior Analysis program at West Virginia University. She is on the editorial board of the *Behavior Analyst.* Her interests include functional assessment and functionally derived interventions, pediatric feeding disorders, and parent training.

Briley Proctor is an Assistant Professor in the Combined Doctoral Program in Counseling Psychology and School Psychology at The Florida State University. She is also the Co-Director for Research and Training of FSU’s Adult Learning Evaluation Center. Her research interests include assessment of intelligence and achievement, academic interventions, and professional issues.