School Psychology Review
1996, Vol. 25, No. 1, pp. 94–107

Situational Specificity and Generality of Test Behaviors for Samples of Normal and Referred Children

Joseph J. Glutting and Eric A. Youngstrom
University of Delaware

Tom Oakland
University of Texas at Austin

Marley W. Watkins
Arizona State University

Abstract: The uses of observations generated during testing were examined through (a) a quantitative synthesis of the available research literature, (b) a study conducted with a nationally representative sample of children (N = 640, including 71 with additional behavioral information), and (c) a study completed with children referred for psychoeducational evaluations (N = 140). Results demonstrated that behavioral and temperament qualities evaluated by test observations are related to children's performance on the Wechsler Intelligence Scale for Children–Third Edition (WISC-III; Wechsler, 1991). Test observations provided considerably less insight into children's adaptation and adjustment outside the test-session environment. IQs from the WISC-III showed limited prediction of classroom behavior, indicating that as much as 97% of the score variation is independent. Findings are examined in light of research on the situational specificity of behavior.

Recording test-session behaviors is an important practice that fulfills a variety of assessment functions. This article describes possible purposes for evaluating test-session behavior and empirically examines three commonly assumed reasons for monitoring such activities: (a) to help illuminate the scorability of children's responses on formal tests administered during the same test session, (b) to document the degree to which the test environment is conducive to optimal performance, and (c) to provide a sample of behavior from a controlled setting that may then generalize and describe children's behaviors and attitudes in other situations. A best evidence synthesis will be used to evaluate these premises, reviewing the results of published studies and test manuals, followed by two empirical studies. The two studies will use a new instrument for evaluating children's test behavior with both a large, nationally stratified sample and a smaller sample of children referred for psychoeducational testing. Results will provide evidence to support the value of formally recording test-session behaviors, as well as indicating limitations on how behavioral observations from test sessions generalize to other contexts.

The observation and recording of test behaviors is routine in the field of individual appraisal. For example, students attending graduate programs in school and clinical psychology are taught to evaluate the behaviors children display when responding to items on individually administered tests. Likewise, textbooks on individual appraisal and intelligence testing encourage the recording and interpreting of children's test behaviors (Kamphaus, 1993; Kaufman, 1994; Sattler, 1988).

Test observations completed by trained clinicians offer certain presumed advantages over the behavior ratings collected by parents and teachers. Specifically, test observations offer a standardized methodology for comparing a child's behavior to the behavior of others. The uniformity of conditions under which test observations occur render them relatively free of environmental variation (e.g., home or classroom climate) that can interfere with the collection of valid observations. None of the major contexts of child development (e.g., home, school, and community) offers as high a level of professional expertise, observational control, or uniformity of conditions as the context of individual test-taking.

The relative objectivity of examiners is another benefit of test observations. Parents, teachers, and others entrusted to rate children's behaviors often lack formal training in data col-

Address all correspondence concerning this article to Joseph J. Glutting, 211 Willard Hall, Department of Educational Studies, University of Delaware, Newark, DE 19716. Electronic mail may be sent via Internet to glutting@strauss.udel.edu.

lection. Parents, in particular, can find observational measures difficult to complete because they may not be familiar with the behavior of children at a given age and grade level. Similarly, extraneous variables (e.g., a desire to have a child either receive or not receive special services) can influence the ratings supplied by parents and teachers. Examiners, on the other hand, are well-versed in the application of observational systems, knowledgeable about child development, and less likely to have a vested interest in a given diagnostic outcome.

Role and Function of Test Observations

Three purposes often are assumed for the collection of test observations. First, test observations help to determine the scorability of responses to test items and serve as cross-checks on the validity of children's formal test scores. The importance and verity of this first purpose is evident because test observations shed light on unusual physical characteristics (e.g., sensory impairments, physical disabilities), and provide an opportunity to document special cultural and educational backgrounds that can adversely impact the validity of scores children obtain on formal standardized tests (see AERA, APA, and NCME Standards, 1985, pp. 73–80).

Second, and related to the first, is the assumption that the interpretation of formal test scores accurately reflects children's abilities and/or achievements. Therefore, a primary goal of individual appraisal is to provide a conducive test environment. Children must be sufficiently at ease with the test situation, motivated, and aware of expectations to perform optimally. Standardized tests such as the Wechsler Intelligence Scale for Children–Third Edition (WISC-III; Wechsler, 1991) and the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983) try to stimulate high levels of motivation by means of attractive packaging and the use of manipulatives and other intriguing materials. However, despite these efforts, Dahlstrom (1993) recently cautioned that, "as important as optimal and appropriate motivation is to the successful execution of any test procedure, it is surprising to find that many published instruments still lack suitable methods of documenting a subject's test-taking compliance" (p. 396). A formal, nomothetic method of recording and evaluating test-session behaviors would provide a mechanism for monitoring compliance, and thus, facilitate judgments about the extent to which children's formal test scores reflect their underlying abilities and achievements.

A third assumed purpose of test observations is to provide a sample of behaviors that may reflect generalizable characteristics. This purpose parallels the primary function of individual testing per se; namely, to discover test-session phenomena that characterize children's growth and development in broad ways. Observations about children's cooperation, persistence, or engagement during testing often are regarded as indicative of propensities in broader life situations, and such observations are commonly believed to be useful for identifying pathogenic factors associated with current behaviors (Kaufman & Reynolds, 1984). If test behaviors reflect dispositional qualities, then formal assessment of such behavior might permit prediction of children's behavior in other, more important contexts. Therefore, examiners are disposed to assume that certain test behaviors are not merely situational, but reveal children's typical and enduring qualities.

Relations Among Test Behavior, Ability, and Behavior in Other Contexts

When the concept of criterion validity is applied to test observations, it concentrates on relationships between children's test-session behavior and their scores on formal tests (e.g., IQ test scores, achievement test scores, etc.). *Intrasession validity* is a term that has been used to describe the strength of associations between measures of test-session behaviors and the formal test scores they accompany (Glutting & McDermott, 1988; Glutting, Oakland, & McDermott, 1989). Thus, intrasession validity addresses the first and second assumed purposes of recording test-session behaviors, because it examines the potential impact of behaviors on children's test scores and indicates the extent to which children's performance can actually be considered optimal.

The construct of ecological validity makes it important to examine relationships between test behaviors evaluated across similar contexts, as well as the generalizability of test behaviors to diverse settings (cf. Neisser, 1991). The term *exosession validity* — similar in meaning to the constructs of external validity or generalizability — has been used to describe how robustly con-

clusions about children's behaviors measured in the context of a particular test session might be related to their behavior and conduct in other situations (Glutting et al., 1989). In other words, exosession validity is determined by evaluating the convergence between recorded test-session behaviors and behaviors observed in other contexts, such as teacher reports, which typically sample a broad array of behaviors over an extended period of time. Thus, the third previously identified purpose of test observations, providing a sample of behavior that can be generalized to other contexts, is a matter of exosession validity.

A related issue also should be considered. Substantial clinical information suggests that low IQ scores are associated with behavior problems in the school or home environments (e.g., Cook, Greenberg, & Kusche, 1994; Kolvin, Miller, Fleeting, & Kolvin, 1988; Tonge & Einfeld, 1991). However, in spite of the importance of this topic to both child development and clinical theory, few studies have systematically examined the degree to which scores on individually administered IQ tests are correlated with children's conduct in other contexts (as rated by parents or teachers). For this reason, a quantitative synthesis of available research findings was carried out on the topics of: (a) the intrasession validity of test observations, (b) the exosession validity of test observations, and (c) relationships between children's IQ and their home or school deportment.

Synthesis of Previous Research

A small, but growing body of literature has begun to examine the validity of test observations. We conducted a best evidence synthesis of previous research as a preliminary matter to the two empirical studies that will be presented in this article. Various procedures were used to locate studies. First, terms such as "test behavior," "test-session behavior," "IQ-behavior," "intelligence behavior," "intelligence-personality," "IQ-conduct disorder," "IQ-attention deficit," "IQ-avoidance," and "IQ-anxiety withdrawal" were used to search the PsychINFO (American Psychological Association, 1991) data base for relevant sources. Sources identified in this way were then examined for references to other relevant studies. The studies included in this synthesis were published in refereed journals in the social sciences or in the technical manuals for psychometric instruments, and were limited to studies with children (not adults) that reported scores on a principle indicator (i.e., full scale IQ, the mental processing composite, or an equivalent — not simply using a subtest as an indicator of overall ability).

Average coefficients were calculated according to procedures recommended by Hunter, Schmidt, and Jackson (1982) and Rosenthal (1991). When a single study provided multiple relevant correlation coefficients, they were averaged to provide a single coefficient per sample. Correlations were then weighted by sample size and averaged across studies as described by Hunter and colleagues (1982). Correlations were not corrected for the reliability of measures because such adjustments would increase their obtained value and not accurately reflect the degree of relationship that researchers and practitioners are likely to encounter in their own work (e.g., Rosenthal, 1991). Confidence intervals also were calculated to gauge whether apparently different correlations actually reflect different degrees of association and to provide some indication of the underlying variability in estimation.

Six studies were found on the topic of intrasession validity.[1] The sources yielded a total of 33 correlation coefficients. The overall relationship was –.34 (95% confidence interval: –.29 to –.41) between children's test behaviors and IQs obtained during the same test session. Four sources were identified for the topic of exosession validity.[2] The studies produced 26 correlations. The average relationship was .18 (95% confidence interval: .15 to .22) between children's test behaviors and their conduct in other contexts, such as in their classroom or community. Lastly, 11 investigations compared associations between children's IQ and their home and school behavior.[3] A total of 38 coefficients was available and the relationship was –.19 (95% confidence interval: –.22 to –.17).

The pattern of coefficients shows modest but meaningful levels of intrasession validity (average $r = -.34$). Moreover, the magnitude of intrasession validity is higher than that found for exosession validity or for relationships between children's IQ and their home and school adjustment, as demonstrated by the fact that the respective 95% confidence intervals do not overlap (even after differences in the directions of the correlations were taken into account by

ignoring signs associated with the coefficients). The results were expected and lead to the inference that test observations possess reasonable intrasession validity, but more limited exosession validity. This inference comports well with information regarding the situational specificity of children's behavior. A meta-analysis by Achenbach, McConaughy, and Howell (1987) demonstrated that much of the behavior observed by parents at home, and teachers in school, is contextually dependent and specific to the situation in which it occurs.

Limitations of Previous Research

Test observations in the studies just cited were obtained through the use of ad hoc item collections or structured evaluations on the Test Behavior Observation Guide (TBOG; Caldwell, 1951; Watson, 1951) or Stanford Binet Observation Schedules (SBOS; Terman & Merrill, 1960; Thorndike, Hagen, & Sattler, 1986). The TBOG and SBOS have long-standing histories in the field of clinical assessment and, like all current measures for rating children's test behaviors — including such instruments as the Behavior and Attitude Checklist (Sattler, 1988) and Test Behavior Checklist (Aylward & MacGruder, 1986) — are of some value in codifying test observations. Nevertheless, a number of qualities limit the utility of these measures.

Sound observation should consider all relevant and verifiable aspects of child functioning, including normal development (McDermott, 1986). Unfortunately, most items on test-behavior scales overlook normal and healthy adjustment. Instead, they limit themselves to evaluating pathological symptoms and negative behaviors. This approach is especially shortsighted because the ability to record healthy and suitable behaviors can be increased simply by altering item valences on rating scales.

A potentially more serious problem relates to the identification of integral dimensions (i.e., scales) underlying items sets. The majority of test behavior instruments are composed of undifferentiated lists of symptoms or inductively derived symptom "domains." Interestingly, the measures used in previous studies present no evidence in support of either a single unifying construct of test behavior or of separate domains of behavior.

Perhaps the greatest deficiency of most current measures of test behavior is the absence of the very basis for making differential child comparisons. Typically, these scales do not supply norms for evaluating how one child's behaviors compare to those of others or for determining the extent of children's compliance to test demands. As a result, examiners are left to their own resources in determining when a given child's test behavior is normal and compliant versus exceptional and noncompliant.

The two empirical studies that follow employ a new, nationally standardized scale of test behavior. The studies explore intrasession validity and exosession validity using the scale's normative cohort and a separate sample of children referred for psychoeducational evaluations. In addition, relationships between children's IQs and their school behavior are analyzed for the referred sample.

Observational Measure

The Guide to the Assessment of Test Session Behavior for the WISC-III and WIAT (GATSB; Glutting & Oakland, 1993) is a brief (29-item), behavior-rating instrument for quickly and reliably evaluating the test-session behavior of children administered one or both of the WISC-III or the Wechsler Individual Achievement Test (WIAT; Wechsler, 1992). Designed for use with children aged 6 years 0 months, to 16 years 11 months, the GATSB was conormed with the WISC-III (N = 640) and a second norming was completed with the WIAT (N = 640). Thus, examiners can use the GATSB to accurately determine the level of children's compliance during administrations of the WISC-III and WIAT and/or evaluate whether a child's behavior differs substantially from the behavior of other children. Because examiners complete ratings (using the 3-point alternatives of "usually applies," "somewhat applies," "doesn't apply") immediately after testing, the process of rating does not interfere with the child's WISC-III or WIAT test performance, and the behavioral picture is recorded while easily recalled.

Principal components analysis and principal axis factor analysis (using both orthogonal and oblique rotations) yielded three dimensions for the standardization sample: Avoidance, Inattentiveness, and Uncooperative Mood. These dimensions are theoretically congruent, align with findings from previous studies of children's test behavior, and are similar to established di-

mensions for evaluating children's adjustment and well-being (cf. Achenbach & Edelbrock, 1983; Quay, 1986). Empirically derived standard scores are offered for each of the three scales and for the total score, which is a combination of scores from the GATSB's scales (*M*s for the four scales = 50, *SD*s = 10). Alpha coefficients for the normative sample show that the GATSB supplies reliable estimates of children's test behaviors (respectively, *r*s = .86, .84, .88, .92).

Method

Samples

Data for the current study included all children from the GATSB's WISC-III standardization sample ($N = 640$). Equal numbers of males and females were included in the sample. Norms for the GATSB are reported according to three age levels. An analysis of variance (ANOVA) was conducted on raw scores to determine whether norms should be provided for each year of age for which the GATSB was designed or whether several year levels could be collapsed into groups without losing any essential information. Results showed three age levels were required: 6–8 years ($n = 212$), 9–12 ($n = 211$), and 13–16 ($n = 217$). Teacher ratings of classroom behavior also were collected for a smaller subset of 71 children from the standardization sample.

Within each age level, children were stratified by race and parent education levels. Four categories of race/ethnicity were employed: White, African American, Hispanic, and Other. Similarly, the sample was stratified according to five parent education levels: 8th grade or less, 9th through 11th grade, high school graduate or equivalent, 1 through 3 years of college or technical school, and 4 or more years of college. Stratification precision for race and parent education levels is within 1 percentage point (plus or minus) of 1988 U.S. Census data. In addition, children's ability levels were required to be normally distributed and parallel those from the WISC-III standardization sample (i.e., Full Scale IQ $M = 100$, $SD = 15$). Otherwise, ratings on the GATSB would have been atypical of the common distribution mean, standard deviation, skewness, and kurtosis of children's ability scores.

The second sample was drawn from children referred for psychoeducational evaluations in the state of Arizona ($N = 140$; 52 females). The children attended public schools and ranged in age from 6 years 0 months to 16 years 0 months ($M = 11.3$ years, $SD = 2.3$ years). Of the sample, 54% were White, 9% were African American, and 37% were Mexican American. All of the children were English dominant.

Criterion Measures

Standardized scales were used as validity criteria. For the normative cohort, intrasession validity was assessed through IQs from the WISC-III: Full Scale IQ (FSIQ), Verbal Scale IQ (VIQ), Performance Scale IQ (PIQ), Verbal Comprehension Index (VCI), Perceptual Organization Index (POI), Freedom from Distractibility Index (FDI), and Processing Speed Index (PSI). Exosession validity consisted of an omnibus behavior-problem score calculated from three teacher-rating scales. Specifically, each child had one of the following three scores: (a) the Total Scale score from the Teacher Referral Form (Achenbach & Edelbrock, 1983), (b) the Total Quotient from the Behavior Evaluation Scale (McCarney, Leigh, & Cornbleet, 1983), or (c) the Problem Behaviors score from the Social Skills Rating System (Gresham & Elliott, 1990). The problem score was simply whichever of the three omnibus indices was available for a child.

Intrasession validity for the referred sample was evaluated by assessing the strength of relation between test-session behaviors (as measured by the GATSB) and performance on the WISC-III as a criterion measure. Exosession validity (i.e., the generalizability of observations from the test session to other settings) was evaluated through teacher ratings on the Adjustment Scales for Children and Adolescents (ASCA; McDermott, 1994). The ASCA was normed on 1,400 5- through 17-year-old children stratified according to 1990 U.S. Census data on the following variables: age, gender, academic level, ethnicity, handicapping condition, national region, community size, and parent education. Standard scores on the ASCA are expressed as *t*-scores ($M = 50$, $SD = 10$). Exploratory and confirmatory components analysis of the standardization sample uncovered six core scales: Attention Deficit-Hyperactivity, Solitary Aggressive (Provocative), Solitary Aggressive (Impulsive), Oppositional Defiant, Dif-

Table 1
Distribution Statistics and Correlations of GATSB Predictors and WISC-III Criteria for the Normative Cohort

	M	*SD*
GATSB predictor		
Total score	50.2	10.1
Avoidance	50.4	10.0
Inattentiveness	49.8	9.8
Uncooperative mood	50.7	9.6
WISC-III criteria		
Full Scale IQ (FSIQ)	100.0	15.2
Verbal Scale IQ (VIQ)	98.8	15.4
Performance Scale IQ (PIQ)	101.7	15.4
Verbal Comprehension Index (VCI)	98.9	15.2
Perceptual Organization Index (POI)	101.3	15.7
Freedom from Distractibility Index (FDI)	101.5	15.1
Processing Speed Index (PSI)	103.4	14.8

	WISC-III criteria						
GATSB predictor	**FSIQ**	**VIQ**	**PIQ**	**VCI**	**POI**	**FDI**	**PSI**
Total score	−.36	−.37	−.31	−.33	−.30	−.30	−.23
Avoidance	−.39	−.37	−.33	−.36	−.31	−.33	−.23
Inattentiveness	−.21	−.20	−.19	−.19	−.17	−.20	−.13
Uncooperative mood	−.28	−.26	−.25	−.24	−.23	−.23	−.20

Note. $N = 640$.

fident, and Avoidant. The ASCA also yields two overall dimensions: Overreactivity (obtained by adding item scores from the first 4 core scales) and Underreactivity (based on item scores from the last 2 core scales). All eight ASCA scores were used as criteria.

Procedures

Examiners conducting the assessments were experienced in the individual administration of ability and achievement tests. All of the examiners were White ($n = 146$ for the normative cohort, $n = 8$ for the referred sample). Test behaviors were assessed through GATSB's completed immediately after administrations of the WISC-III. Classroom teachers evaluated children on the behavior rating scales within a month of the test sessions.

Results

Scores on the WISC-III show a strong concordance with the distribution of children's ability levels in the population. Distributional statistics for the GATSB and WISC-III normative cohort are presented in the upper part of Table 1. The lower part of Table 1 provides intrasession validity coefficients between *t*-scores from the GATSB's four scales and IQs from the WISC-III. All 28 of the correlation coefficients are statistically significant ($p < .0010$, which could be compared with a Bonferroni-adjusted critical value of alpha = .0018 to maintain an overall alpha level of .05). The coefficients reveal that, among the GATSB's three primary scales, Avoidance demonstrates the highest general relationship to children's ability levels. Uncooperative Mood has the second highest relationship and is followed by Inattentiveness.

The average correlation is −.27 (with 95% confidence interval of −.34 to −.17) between scores from the GATSB's four scales (including the total score) and IQs from the WISC-III. This averaged coefficient is congruent with the intrasession validity found in the meta-analysis of previous studies (average $r = -.34$) and it supports inferences that test observations show modest but meaningful levels of intrasession validity.

Table 2
Differences in WISC-III Full Scale IQs for Children Showing Compliant and Noncompliant Test Behaviors

	Mean WISC-III FSIQ	*t*-value	*p*
GATSB Total score			
Compliant (n = 547)	101.5		
Noncompliant (n = 93)	91.1	6.31	.001
GATSB Avoidance scale			
Compliant (n = 536)	101.9		
Noncompliant (n = 104)	90.5	7.27	.001
GATSB Inattentiveness scale			
Compliant (n = 551)	101.1		
Noncompliant (n = 89)	93.7	4.26	.001
GATSB Uncooperative Mood scale			
Compliant (n = 559)	101.3		
Noncompliant (n = 81)	91.6	5.49	.001

Although the relationships just presented between test-session behavior and ability are significant (i.e., average $r = -.27$, average $p < .001$), statistical significance does not speak to the pragmatic value (i.e., clinical or psychological significance) of the associations. We examined the question of practical utility by dividing children from the GATSB normative sample according to whether they exhibited compliant or noncompliant test behaviors during administrations of the WISC-III. Comparison groups were formed using GATSB *t*-scores in the average range (i.e., ≤ 59) versus *t*-scores one standard deviation above the mean (i.e., ≥ 60). Results (presented in Table 2) show that children with compliant test behaviors on the GATSB earn, on average, WISC-III FSIQs 7 to 10 points higher than those with noncompliant behaviors. The differences are striking and show that children with compliant test behaviors earn WISC-III IQs between one-half to two-thirds of a standard deviation higher than children with less suitable test behaviors. Thus, it becomes apparent that children's test behaviors are meaningfully related to the magnitude of IQs they obtain on the WISC-III.

Intrasession validity also was examined for the referred sample (see Table 3). As expected, IQs on the WISC-III are lower (M FSIQ = 81.7). It is important to note that standard deviations obtained with the referred sample are generally comparable to what would be expected in other samples (see Tables 3 and 4). Thus, it appears that range restrictions, commonly a problem when working with referred samples, are not operating to attenuate the magnitude of correlations observed in this particular sample.

The pattern of associations among the GATSB's three primary scales and the WISC-III criteria are somewhat different for the referred sample than those reported earlier for the normative group. Avoidance continues to show the highest set of associations with children's ability levels, whereas the pattern reverses for the two other GATSB scales: Inattentiveness now has the second highest connection and is followed by Uncooperative Mood. The average overlap is $-.24$ across the 28 coefficients. This mean is somewhat lower than levels just reported for the GATSB's normative cohort (average $r = -.27$) and from previous studies (average $r = -.34$). Nevertheless, it remains higher than the exosession validity obtained during earlier studies (average $r = .18$) and it supports contentions that test observations are modestly, but meaningfully, related to the IQs of children referred for psychoeducational evaluations.

Exosession validity between children's test behaviors and their classroom conduct was evaluated for only a small segment the GATSB's normative group ($n = 71$). Another shortcoming is that comparisons were confined to a single

Table 3
Distribution Statistics and Correlations of GATSB Predictors and WISC-III Criteria for a Referred Sample

	M	*SD*
GATSB predictor		
Total score	52.6	8.7
Avoidance	54.4	10.3
Inattentiveness	50.8	8.2
Uncooperative mood	50.7	7.5
WISC-III criteria		
Full Scale IQ (FSIQ)	81.7	15.3
Verbal Scale IQ (VIQ)	80.6	15.0
Performance Scale IQ (PIQ)	86.2	15.6
Verbal Comprehension Index (VCI)	81.4	15.2
Perceptual Organization Index (POI)	87.1	16.7
Freedom from Distractibility Index (FDI)	82.0	13.9
Processing Speed Index (PSI)[1]	87.5	15.1

	WISC-III criteria						
GATSB predictor	**FSIQ**	**VIQ**	**PIQ**	**VCI**	**POI**	**FDI**	**PSI**
Total score	−.33	−.28	−.34	−.25	−.32	−.31	−.31
Avoidance	−.39	−.33	−.39	−.33	−.38	−.29	−.23
Inattentiveness	−.12	−.10	−.13	−.07	−.10	−.21	−.28
Uncooperative mood	−.17	−.15	−.18	−.11	−.13	−.18	−.21

Note. $N = 140$.
[1]The Symbol Search subtest which contributes to the WISC-III PSI was not administered to all children. Consequently, $n = 96$.

omnibus score of classroom behavior compiled from three instruments. The associations follow: .12 between the GATSB total score and the omnibus score of children's classroom behavior, .22 for the Avoidance scale, .04 for Inattentiveness, and .17 for Uncooperative Mood.

The analysis was repeated with the referred sample ($N = 140$). The second analysis compensated for shortcomings of the first exosession validity study. The teachers of children in the referred group used the same rating scale (i.e., the ASCA). Furthermore, rather than using a solitary reckoning of classroom adjustment, comparisons were directed to associations between the GATSB's 4 scores and each of ASCA's 8 measures.

The GATSB and ASCA evaluate similar constructs, but do so in different contexts. Strong construct validity is suggested whenever an appropriate pattern of convergent and divergent associations is found between similar tests (Campbell, 1960; Thorndike, 1982). Higher correlations were expected between identical, or convergent, scales of the GATSB and ASCA (e.g., GATSB Avoidance and ASCA Avoidance) and lower correlations were expected between divergent scales (e.g., GATSB Avoidance and ASCA Attention Deficit-Hyperactive).

Table 4 presents exosession validity coefficients between the GATSB and ASCA. Results show that the anticipated convergent associations are collectively higher than divergent associations. Nonetheless, even though the pattern of coefficients supports suppositions of construct validity, the magnitude of the relations is trivial. Only 13 of the 32 correlations are statistically significant ($p < .05$) even without adjusting the alpha level for the total number of comparisons made. More important, the mean coefficient is quite low (average $r = .16$, with a 95% confidence interval of −.01 to +.32) and indicates that approximately 97% of the variation in scores on either type of measure is unique (i.e., $1-.16^2 = 97.4$). The current outcomes between

Table 4
Distribution Statistics and Correlations of GATSB Predictors and ASCA Criteria for a Referred Sample

	M	*SD*
GATSB predictor		
Total score	52.6	8.7
Avoidance	54.4	10.3
Inattentiveness	50.8	8.2
Uncooperative mood	50.7	7.5
ASCA criteria		
Avoidant (AV)	53.4	10.8
Diffident (DIF)	52.3	11.3
Underreactivity(Under)	54.2	10.4
Solitary Aggressive–Provocative (SAP)	55.1	13.1
Solitary Aggressive–Impulsive (SAI)	55.3	12.4
Attention Deficit-Hyperactivity (ADH)	58.0	10.5
Oppositional Defiant (OD)	57.6	16.2
Overreactivity (Over)	59.2	11.1

	ASCA criteria							
GATSB predictor	**AV**	**DIF**	**Under**	**SAP**	**SAI**	**ADH**	**OD**	**Over**
Avoidance	**.23**	**.37**	**.39**	.05	.06	.01	.02	.01
Inattentativeness	.13	.00	.03	**.20**	**.33**	**.24**	**.17**	**.24**
Uncooperative mood	.10	.10	.05	**.19**	**.29**	**.21**	**.05**	**.20**
Total score[1]	.20	.23	.24	.15	.23	.17	.10	.17

Note. $N = 140$. Convergent associations are presented in boldface.

[1]Because the total score is a composite calculated from all of the GATSB's primary scales, it was not used to analyze patterns of convergent and divergent validity.

the GATSB and ASCA are comparable to previous findings of exosession validity (average $r = .18$) and argue for the conservative generalization of children's test behaviors to other contexts.

A final analysis examined connections between IQs from the WISC-III and classroom behaviors measured by the ASCA. The most interesting comparisons are likely to be those for the third and fourth factors of the WISC-III. Low scores on the WISC-III's Freedom from Distractibility Index (FDI) have been linked to a variety of personality problems, including: inattention, distractibility, hyperactivity, poor study skills, somatic complaints, and acting-out behaviors (see Wielkiewicz, 1990, for a review). Similarly, the WISC-III's new, fourth Processing Speed Index (PSI) has been shown to be a sensitive discriminator of children with attention deficit-hyperactivity disorder (Prifitera & Dersh, 1993; Schwean, Saklofske, Yackulic, & Quinn, 1993), as well as for those evidencing severe emotional disturbance (Teeter & Smith, 1993).

Four of the 56 correlations between the WISC-III and ASCA reached statistical significance, at $p < .05$ (see Table 5). This ratio barely exceeds the chance rate of 3 significant connections (i.e., $56 \times .05 = 2.8$). The average coefficient also is meager (average $r = -.04$, with a 95% confidence interval of $-.20$ to $+.13$) and indicates that as much as 99% of the score variation is unique to each instrument. Regarding the FDI and PSI, their typical relationship is $-.11$ with classroom adjustment variables measured by the ASCA. The highest correlation of the WISC-III's third and fourth factors occurs between the PSI and Oppositional Defiant classroom behaviors ($r = -.27$). However, even this strongest of relationships has at least 93% unique score variation. The current findings of meager overlap should come as no surprise: 11

Table 5
Correlations Between WISC-III IQs and ASCA Ratings for a Referred Sample

WISC-III predictor	ASCA criteria							
	AV	DIF	Under	SAP	SAI	ADH	OD	Over
FSIQ	−.14	−.16	−.16	.05	.14	.03	.05	.04
VIQ	−.09	−.12	−.09	.05	.14	.03	.06	.06
PIQ	−.15	−.19	−.21	.05	.13	.03	.05	.02
VCI	−.06	−.12	−.07	.05	.15	.06	.06	.08
POI	−.13	−.15	−.17	.08	.18	.06	.07	.06
FDI	−.13	−.04	−.05	.07	.08	−.05	.06	.00
PSI[1]	−.17	−.18	−.22	−.22	−.19	−.19	−.27	−.26

Note. $N = 140$. AV = Avoidant; DIF = Diffident; Under = Underactivity; SAP = Solitary Aggressive–Provocative; SAI = Solitary Aggressive–Impulsive; ADH = Attention Deficit-Hyperactivity; OD = Opposition Defiant; Over = Overreactivity; FSIQ = Full Scale IQ; VIQ = Verbal Scale IQ; PIQ = Performance Scale IQ; VCI = Verbal Comprehension Index; POI = Perceptual Organization Index; FDI = Freedom from Distractibility Index; PSI = Processing Speed Index.

[1]The Symbol Search subtest, which contributes to the WISC-III PSI, was not administered to all children. Consequently, $n = 96$.

previous investigations showed an average relationship of −.19 between children's scores on individually administered IQ tests and their home and school behavior.

Discussion

Results from the following three sources converge to indicate that test observations possess modest, but meaningful, levels of intrasession validity: a quantitative synthesis of prior research, a correlational analysis of data from a nationally representative sample, and another correlational analysis of data from a sample of referred children. The analyses also showed that children with noncompliant behaviors, as measured by the GATSB, were likely to obtain WISC-III FSIQs anywhere from 7 to 10 points lower than children with more suitable test behaviors. Effect sizes this large (.5 to .7, and/or more than one-half of a standard deviation between groups) represent a substantial difference in IQs (cf. Cooper, 1989). These results consistently demonstrate that children's test behaviors are meaningfully related to the magnitude of scores they obtain on IQ tests.

The appreciable IQ differences found between compliant and noncompliant children may cause some clinicians to assume that test behaviors are causal components of ability scores. It must be emphasized that the relations are correlative and in no way imply causation. Rather, children with noncompliant test behaviors are simply more likely to obtain lower IQs than compliant children on average. At the same time, the magnitude of the IQ differences speaks to the importance of observing behaviors peripheral to scorable test responses, and it highlights the need for psychologists to faithfully record children's behavioral dispositions during testing.

The generalizability of test behaviors across situations can be assessed by their associations with measures of children's adjustment in important contexts of development. The present results argue for either the extremely conservative generalization of test behaviors or for no extrapolation at all. Notwithstanding several statistically significant correlations, current findings of both a nationally representative and a referred sample indicate that approximately 97% of the variation in test behaviors is specific to the context in which they occur. These findings are remarkably similar to results obtained from a quantitative synthesis of earlier test-behavior studies and parallel literature addressing the generality of behaviors across home and school environments (Achenbach et al., 1987).

Individualized testing is a unique activity, distinct from everyday contexts. Limited exosession validity can be expected whenever observation time is reduced and the sampling of behavior variation is constrained, which is the typical situation for test-session observations. Of interest here is the relative lack of support for popular claims that test-session behaviors

are generalizable to important phenomena found in natural child environments, such as behavioral adjustment and competence at home or in school. These findings also are consistent with a more ecological perspective: the constraints and demand characteristics of the individual testing environment are likely to elicit a sample of behaviors with relatively low associations to the same child's behavior in other settings.

The situational specificity of test behaviors may actually enhance their utility. Schachar, Rutter, and Smith (1981) demonstrated that only a small portion of children with attention deficit-hyperactivity disorder (ADHD) displayed symptoms across both parent and teacher ratings. Children with cross-environment ADHD showed more noteworthy academic and cognitive impairments than children with situational hyperactivity. The situational specificity of children's test behaviors may offer proof that test observations of poor sustained attention, deficient impulse control, and excessive activity level will compliment or clarify behavioral descriptions provided by parents and teachers, and in this way, increase diagnostic precision.

As for the usefulness of IQs in predicting behavior, results showed marginal overlap between IQs from the WISC-III and children's classroom demeanor on the ASCA. This last set of findings indicate that children are not well-served by behavior hypotheses generated from the score patterns and IQs they receive on individually administered tests of ability (e.g., inferences associated with inattention, distractibility, or somatic complaints). Instead, observations obtained from parent and teacher ratings are better methods for assessing children's emotional adjustment because their utility is well-documented and supported empirically.

Future Directions

The results of the present studies could be expanded in a variety of ways. The quantitative synthesis of findings in the literature could be extended to include more sources, making feasible the identification of mediating variables (or sources of heterogeneity) that might affect relations between test-session behaviors and either IQ or more general reports of behavior. The goal of the present synthesis was to combine the effects reported in sources of good quality and to establish a baseline against which to judge the outcomes of our empirical studies. However, whether or not the strength of association between the constructs of test-session behavior, IQ, and behavior across other settings is uniform across differing samples and research designs is a question worth further investigation. Similarly, although the present study relied on an instrument with well-established validity, reliability, and utility, it would be helpful to have additional studies examine test-session behavior with other measures of IQ besides the WISC-III. Finally, the potential correlation between test-session behavior and stable individual constructs such as personality and temperament has not been directly evaluated, although the relative lack of generalizability of test-session behaviors to behavior in other contexts suggests that such a relation would account for only a small portion of the variance in either construct.

Additional pragmatic reasons for considering test-session behavior in a formalized way include concerns of professional liability and the potential for confounding of scores by halo effect. School psychologists must be alert to legal issues, including possible litigation that might ensue from (a) failing to follow established protocols of test administration, (b) improper scoring and tabulation of test results, and (c) creating an atmosphere detrimental to optimal test performance. The established intrasession validity of test observations makes them useful for documenting whether important test behaviors affect the quality of scores children obtain on standardized tests. The formal documentation of test-session behaviors not only helps assess the interpretability of scores, but also records the extent to which the testing situation is conducive to optimal performance.

Psychologists conducting individual appraisals have direct knowledge of children's ability levels. Halo effects occur when psychologists are positively or negatively disposed by children's IQ-test scores and rate test behaviors accordingly. An assessment of qualities directly measured by a test and those related to test-taking behaviors are not independent. It is likely that ratings of test-taking behaviors are influenced to some unknown (and perhaps unknowable) degree by children's overall intellectual performance, and attempts to separate the two would not be possible during traditional one-to-one clinical assessments.

The distorting effects of halo are an impediment to accurate appraisal. Even so, no evidence exists that knowledge of children's test performance and their test-taking behaviors invalidates or confounds one or both measures. Furthermore, the professional training of school psychologists, the uniformity and observational control of test sessions, and the lack of a vested interest in a diagnostic outcome make test observations appear less susceptible to halo than behavior ratings completed by parents and teachers.

The correlational studies in this article used the GATSB, a new measure for evaluating children's test behaviors. The studies uncovered two valid reasons for recording test behaviors: (a) to reliably determine whether children's test-session behaviors are substantially different from those of same-aged peers and (b) to determine the extent to which children's test behaviors affect the quality of scores they obtain on formal IQ and achievement tests. The vast majority of children referred for psychoeducational assessments will display appropriate levels of involvement, attentiveness, and cooperation during test sessions. However, a number of children show inappropriate behaviors. The GATSB is currently the only test-observation measure to provide norms — an essential requirement for any instrument designed to facilitate interindividual comparisons and decision making. Thus, the GATSB has a considerable advantage in identifying children whose test behaviors are unsuitable or inappropriate, and in turn, it will help to describe and label the specific test-behavior dimensions that influence the scores children obtain on formal tests of ability and achievement.

References

Achenbach, T. M., & Edelbrock, C. (1983). *Child Behavior Profile.* Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for education and psychological testing.* Washington, DC: Author.

American Psychological Association. (1991). *PsychINFO psychological abstracts information services users reference manual.* Washington, DC: Author.

Aylward, G. P., & MacGruder, R. W. (1986). *Test Behavior Checklist.* Brandon, CT: Clinical Psychology Publishing.

Beck, B. L., & Spruill, J. (1987). External validation of the cognitive triad of the Personality Inventory for Children: Cautions on interpretation. *Journal of Consulting and Clinical Psychology, 55,* 441–443.

Caldwell, B. M. (1951). Test Behavior Observation Guide. In R. Watson (Ed.), *The clinical method in psychology* (pp. 67–71). New York: Harper & Brothers.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15,* 546–553.

Cook, E. T., Greenberg, M. T., & Kusche, C. A. (1994). The relations between emotional understanding, intellectual functioning, and disruptive behavior problems in elementary-school-aged children. *Journal of Abnormal Child Psychology, 22,* 205–219.

Cooper, H. M. (1989). *Integrating research: A guide for literature reviews.* Thousand Oaks, CA: Sage.

Dahlstrom, W. G. (1993). Tests: Small samples, large consequence. *American Psychologist, 48,* 393–399.

Glutting, J. J., & McDermott, P. A. (1988). Generality of test-session observations to kindergartners' classroom behavior. *Journal of Abnormal Child Psychology, 16,* 527–537.

Glutting, J., & Oakland, T. (1993). *GATSB: Guide to the assessment of test session behavior for the WISC-III and WIAT.* San Antonio, TX: Psychological Corporation.

Glutting, J. J., Oakland, T., & McDermott, P. A. (1989). Observing child behavior during testing: Constructs, validity, and situational generality. *Journal of School Psychology, 27,* 155–164.

Gordon, M., DiNiro, D., Mettelman, B. B., & Tallmadge, J. (1989). Observation of test behavior, quantitative scores, and teacher ratings. *Journal of Psychoeducational Assessment, 7,* 141–147.

Gresham, F., & Elliott, S. N. (1990). *Social Skills Rating System.* Circle Pines, MN: American Guidance Service.

Hinshaw, S. P., Morrison, D. C., Carte, E. T., & Cornsweet, C. (1987). Factorial dimensions of the Revised Behavior Problem Checklist: Replication and validation within a kindergarten sample. *Journal of Abnormal Child Psychology, 15,* 309–327.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Jorm, A. F., Share, D. L., Matthews, R., & Maclean, R. (1986). Behaviour problems in specific reading retarded and general reading backward children: A longitudinal study. *Journal of Child Psychology and Psychiatry, 27,* 33–43.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: Wiley Interscience.

Kaufman, A. S., & Kaufman, N. L. (1983). *K-ABC: Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Reynolds, C. R. (1984). Intellectual and academic achievement tests. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 195–220). New York: Pergamon Press.

Kolvin, L., Miller, F. J. W., Fleeting, M., & Kolvin, P. A. (1988). Risk/protective factors for offending with particular reference to deprivation. In M. Rutter (Ed.), *Studies of psychological risk: The power of longitudinal data* (pp. 77–95). New York: Cambridge UP.

Laosa, L. M. (1986). *Test-taking styles of young children: Longitudinal analyses.*Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Lefkowitz, M. M., & Tesiny, E. P. (1985). Depression in children: Prevalence and correlates. *Journal of Consulting and Clinical Psychology, 53,* 647–656.

Lynam, D., Moffitt, T., & Stouthamer-Loeber, M. (1993). Explaining the relation between IQ and delinquency: Class, race, test motivation, school failure, or self-control? *Journal of Abnormal Psychology, 102,* 187–196.

Matheny, A. P., Brown-Dolan, A., & Wilson, R. S. (1974). Bayley's infant behavior record: Relations between behaviors and mental test scores. *Developmental Psychology, 10,* 696–702.

McCarney, S. B., Leigh, J. E., & Cornbleet, J. A. (1983). *The behavior evaluation scale.* Austin, TX: Pro-Ed.

McDermott, P. A. (1986). The observation and classification of exceptional child behavior. In R. T. Brown & C. R. Reynolds (Eds.), *Psychological perspectives on childhood exceptionality: A handbook* (pp. 136–180). New York: Wiley Interscience.

McDermott, P. A. (1994). *Manual of Adjustment Scales for Children and Adolescents.* Philadelphia: Edumetric and Clinical Science.

McGee, R., Anderson, J., Williams, S., & Silva, P. A. (1986). Cognitive correlates of depressive symptoms in 11-year-old children. *Journal of Abnormal Child Psychology, 14,* 517–524.

McGee, R., Williams, S., & Silva, P. A. (1985). Factor structure and correlates of ratings of inattention, hyperactivity, and antisocial behavior in a large sample of 9-year-old children from the general population. *Journal of Consulting and Clinical Psychology, 53,* 480–490.

Milich, R., Loney, J., & Landau, S. (1982). Independent dimensions of hyperactivity and aggression: A validation with playroom observation data. *Journal of Abnormal Psychology, 91,* 183–198.

Moriarty, A. (1961). Coping patterns of preschool children in response to intelligence test demands. *Genetic Psychology Monographs, 64,* 3–127.

Neisser, U. (1991). A case of misplaced nostalgia. *American Psychologist, 46,* 34–36.

Oakland, T., & Glutting, J. J. (1990). Examiner observations of children's WISC-R test-related behaviors: Possible socioeconomic status, race, and gender effects. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 86–90.

Prifitera, A., & Dersh, J. (1993). Base rates of WISC-III diagnostic subtest patterns among normal, learning-disabled, and ADHD samples. *Journal of Psychoeducational Assessment, WISC-III Monographs,* 43–55.

Quay, H. C. (1986). Classification. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (3rd ed., pp. 1–34). New York: Wiley.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Schachar, R., Rutter, M., & Smith, A. (1981). The characteristics of situationally and pervasively hyperactive children: Implications for syndrome definition. *Journal of Child Psychology and Psychiatry, 1,* 241–247.

Schwean, V. L., Saklofske, D. H., Yackulic, R. A., & Quinn, D. (1993). WISC-III performance of ADHD children. *Journal of Psychoeducational Assessment, WISC-III Monograph,* 56–70.

Teeter, P. A., & Smith, P. L. (1993). WISC-III and WJ-R: Predictive and discriminant validity for student with severe emotional disturbance. *Journal of Psychoeducational Assessment, WISC-III Monograph,* 114–124.

Terman, L. M., & Merrill, M. A. (1960). *Stanford-Binet Intelligence Scale: Manual for the third revision Form L–M.* Boston: Houghton Mifflin.

Tesiny, E. P., Lefkowitz, M. M., & Gordon, N. H. (1980). Childhood depression, locus of control, and school achievement. *Journal of Educational Psychology, 72,* 506–510.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Thorndike, R. L., Hagen, E. P., & Sattler, J. J. (1986). *Stanford-Binet Intelligence Scale: Fourth Edition.* Chicago: Riverside.

Tonge, B., & Einfeld, S. (1991). Intellectual disability and psychopathology in Australian children. *Australia and New Zealand Journal of Developmental Disabilities, 17,* 155–167.

Watson, R. (1951). *The clinical method in psychology.* New York: Harper and Brothers.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children: Third edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1992). *Wechsler Individual Achievement Test.* San Antonio, TX: The Psychological Corporation.

Wielkiewicz, R. M. (1990). Interpreting low scores on the WISC-R third factor: It's more than distractibility. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 91–97.

Footnotes

[1]The following investigations examined intrasession validity: Glutting, Oakland, and McDermott (1989); Gordon, DiNiro, Mettelman, and Tallmadge (1989); Laosa (1986); Lynam, Moffitt, and Stouthamer-Loeber (1993); Matheny, Brown-Dolan, and Wilson (1974); and Moriarty (1961).

[2]The following investigations examined exosession validity: Glutting and McDermott (1988); Glutting et al. (1989); Gordon et al. (1989); and Lynam et al. (1993).

[3]The following investigations examined relationships between children's IQ-test scores and their home and school behavior: Beck and Spruill (1987); Jorm, Share, Matthews, and Maclean (1986); Hinshaw, Morrison, Carte, and Cornsweet (1987); Lefkowitz and Tesiny (1985); McGee, Anderson, Williams, and Silva (1986); McGee, Williams, and Silva (1985); Milich, Loney, and Landau (1982); Oakland (1980); and Tesiny, Lefkowitz, and Gordon (1980).

Joseph J. Glutting, PhD, is an Associate Professor in School Psychology at the University of Delaware. His research interests include the interpretation of results from individually-administered tests of ability, achievement, and personality, and the assessment of children's test-session behavior.

Eric A. Youngstrom, BA, is in the clinical psychology doctoral program at the University of Delaware. His research interests include emotional development and its relation to psychopathology, as well as individual assessment.

Thomas Oakland, PhD, is Department Chair, Foundations of Education at the University of Florida, Gainesville, He is a Fellow of the American Psychological Association, and his research interests span a variety of areas including individual appraisal and children's temperaments. He is senior author of the Student Styles Questionnaire (SSQ).

Marley W. Watkins, PhD, is an Associate Professor in School Psychology at Pennsylvania State University. He is a Diplomate in School Psychology, American Board of Professional Psychology, and his research interests include psychodiagnostic assessment, development of microcomputer software for test interpretations and computer assisted instruction.

This article was accepted during the editorship of Edward S. Shapiro, PhD.