Journal of Clinical Child Psychology
2002, Vol. 31, No. 2, 263–271

The Guide to the Assessment of Test Session Behavior: Validity in Relation to Cognitive Testing and Parent-Reported Behavior Problems in a Clinical Sample

Eric Daleiden
The University of Tulsa

Ronald S. Drabman
The University of Mississippi Medical Center

Jennifer Benton
The University of Tulsa

Examined the validity of the Guide to the Assessment of Test Session Behavior (GATSB) in a clinical sample. The GATSB is a structured and standardized measure that was normed on the Wechsler Intelligence Scale for Children–Third Edition (WISC–III) standardization sample. The purpose of this study was to extend validity research on the GATSB to cognitive measures other than the WISC–III and to parent-reported child behavior problems. Test observations were taken for 122 children ages 6 to 16 years who were referred to an outpatient psychology clinic for psychoeducational evaluations. GATSB scores demonstrated a moderate relation with general indexes from the WISC–III, the Woodcock–Johnson Psychoeducational Battery–Revised (WJ–R), and the Wide Range Assessment of Memory and Learning (WRAML), but displayed low magnitude correlations with the Child Behavior Checklist (CBCL). These results expand existing data supporting the validity of the GATSB and suggest that it may be fruitfully adopted into a clinic setting.

Observing child behavior during psychological assessments can be very helpful in determining the validity of an assessment for that child (Glutting & McDermott, 1988; Glutting, Oakland, & McDermott, 1989). If a child is distracted or uncooperative, then testing may not accurately reflect the child's true abilities. In practice, behavioral observations during assessment are often informal, with the assessor making notes about the child's behavior. However, with the introduction of the Guide to the Assessment of Test Session Behavior (GATSB; Glutting & Oakland, 1993), a brief and easy-to-use standardized instrument is now available to formally assess behavior during a testing session.

The GATSB is a structured, clinician-report measure that was normed on the Wechsler Intelligence Scale for Children–Third Edition (WISC–III; Wechsler, 1993) and the Wechsler Individual Achievement Test (WIAT; Psychological Corporation, 1992) standardization samples. The GATSB is a 29-item questionnaire that is completed by the clinician immediately following a test session. GATSB items gauge the extent to which children respond positively to attempts to establish rapport, persist on difficult tasks, attentively listen to directions and test items, and show interest in test activities. Scoring of the GATSB yields an overall problem score, plus three subscale indexes of Avoidance, Inattentiveness, and Uncooperative Mood.

Efforts to validate the GATSB have examined two primary lines of criterion evidence. First, to validate the use of the GATSB for moderating clinical interpretations of the psychoeducational tests with which it was administered, studies of intrasession validity have examined the correlation between GATSB scores and the IQ or achievement tests that provided the behavior sample for the clinician's judgments. Second, studies of exosession validity have examined the correlation between the GATSB and other measures of child behavior in different settings (e.g., the classroom and home environments) to validate the use of the GATSB for making inferences about the child's likely behavior in non-testing situations. For example, the intrasession validity of the GATSB might be evidenced by correlations between GATSB scores and WISC–III scores, whereas the exosession validity of the GATSB might be evidenced by correlations between GATSB scores and parent-reported Child Behavior Checklist (CBCL; Achenbach, 1991) scores.

This project was supported in part by a Faculty Research Grant (1016183) from the University of Tulsa.

We are grateful for the assistance of Regina Carroll, Sherri Liles, and Elizabeth Neuhaus in performing this project.

Requests for reprints should be sent to Eric Daleiden, Department of Psychology, University of Tulsa, 600 S. College Avenue, Tulsa, OK 74104. E-mail: eric-daleiden@utulsa.edu

To date, studies have provided considerable support for the intrasession validity of the GATSB and mild support for its exosession validity. Studies of the intrasession validity of the GATSB have reported significant correlations (often in the –.30 to –.40 range) between GATSB scores and indexes from the WISC–III and the WIAT (Glutting & Oakland, 1993; Konald, Maller, & Glutting, 1998; Maller, Konald, & Glutting, 1998). A study that examined potential test bias of the GATSB related to sex, ethnicity, and socioeconomic status found that children who exhibited higher levels of avoidance, inattentiveness, and uncooperative behavior tended to exhibit lower WISC–III and WIAT scores (Glutting, Oakland, & Konald, 1994). In comparing various ethnic groups (i.e., Caucasian, African American, and Latino American), the only ethnicity difference on the GATSB was found between Latinos and Caucasians. Latinos with below-average IQs tended to display better test session behaviors than Caucasians with below-average IQs. Nevertheless, the GATSB generally demonstrated similar intrasession validities for children of different race or ethnicity, sex, and socioeconomic status (Konald, Glutting, Oakland, & O'Donnell, 1995).

Evidence for the exosession validity of the GATSB has been observed in the form of correlations on the order of –.15 to –.20 between GATSB scores and teacher-reported or direct observations of classroom behavior (Glutting & Oakland, 1993). Additional evidence has indicated that the GATSB scales, particularly inattentiveness, may be useful in distinguishing children with attention deficit hyperactivity disorder from matched controls (Glutting, Robins, & de Lancey, 1997). Nevertheless, these exosession validity data have prompted the authors to caution against inferring that test session behavior patterns generalize to non-test settings.

These findings highlight the promise of the GATSB as a measure of test session behavior but also reveal directions for continuing the validation process. Examination of exosession validity has emphasized classroom-based and teacher-reported behavior and could be furthered through investigation of home-based and parent-reported behavior. Similarly, intrasession validity studies have supported the GATSB's convergence with the WISC–III and WIAT. To date, studies have not examined whether the GATSB scores obtained based on WISC–III or WIAT administrations can predict child performance on other cognitive measures administered as part of a test battery.

The purpose of this study was to extend research on the GATSB by examining the validity with cognitive measures other than the WISC–III and WIAT and by examining exosession validity in relation to parent-reported behavior problems. Therefore, after completing the GATSB based on administration of the WISC–III, clinicians administered the Woodcock–Johnson Psychoeducational Battery–Revised (WJ–R; Woodcock & Johnson, 1989), and the Wide Range Assessment of Memory and Learning (WRAML; Adams & Sheslow, 1990). The CBCL was administered as an exosession validity measure of parent-report child behavior. We predicted that the GATSB would be negatively correlated with the WJ–R and the WRAML, as it is with the WISC–III. We also predicted that the GATSB should be positively correlated with the CBCL behavior problem measures. Further, we expected a specific pattern of GATSB and CBCL subscale correlations. The CBCL Internalizing Problems and its narrowband scales (i.e., Withdrawn, Somatic Complaints, and Anxious/Depressed) should be positively related with the GATSB Avoidance subscales. Glutting and Oakland (1993) suggested that the GATSB Inattention subscale should be positively related with the externalizing problems. However, this prediction was based on the inclusion of attention problems and attention deficit hyperactivity disorder (Glutting et al., 1997) as externalizing problems. Because attention problems are not included on the Externalizing broadband scale of the CBCL, we predicted that CBCL Externalizing and its narrowband scales (i.e., Delinquent and Aggressive Behavior) would positively correlate with GATSB Uncooperative Mood. The CBCL Attention Problem scale was expected to correlate with GATSB Inattention.

Method

Participants

Participants were 122 children (49 girls and 73 boys) between 6 and 16 years of age ($M = 10.2$, $SD = 3.2$) who received psychoeducational evaluations at an outpatient child psychology clinic. Youth were referred to the clinic for a variety of school and social concerns. The 122 participants represented 98% of 125 consecutive 6- to 16 year-olds who were tested at the clinic during the study period.[1] The sample was predominantly Caucasian and middle to upper-middle class.[2] This sample was representative of the typical assessment clinic clientele, which tended to be more affluent with less minority representation than the broader clinic treatment and general

[1]Three cases were excluded from this sample. First, the GATSB was inadvertently not administered to an 11-year-old boy with a CBCL total problem *T* score of 36 and IQ scores of 88 on the WJ–R and 74 on the WISC–III. Second, a 16-year-old boy who received GATSB *T* scores ranging from 82 to 99 and IQ scores of 49 on the WJ–R and 46 on the WISC–III was excluded due to a missing CBCL. Finally, a 14-year-old girl was excluded due to an extremely low IQ of 22 on the WJ–R (52 on the WISC–III), with GATSB *T* scores ranging from 90 to 99 and CBCL total problems *T* score of 68. If the two cases with low IQ scores and elevated GATSB scores were included in analyses, validity coefficients for the GATSB would be slightly higher than reported.

[2]Race and socioeconomic status were not routinely coded into the clinic database at the individual level, so the specific distribution of these variables in the final sample could not be reported.

catchment area populations. The final sample was of average intellectual ability yet ranged from the mild mental retardation to superior range and scored approximately one standard deviation above the normative mean on various indexes of emotional and behavioral problems (see Table 1). Table 2 presents information on the number of participants displaying various cognitive and behavioral problems based on the evaluation. These data do not represent clinical diagnoses, but rather descriptions of sample characteristics based on the assessment battery results. Of the 122 cases, 30% of the sample demonstrated no cognitive or behavioral problems ($n = 37$), 12% displayed only one pure cognitive problem ($n = 15$), another 12% displayed only one pure behavioral problem ($n = 15$), and 45% displayed comorbid cognitive or behavioral problems ($n = 55$). Of the 55 comorbid cases, 5 displayed only cognitive problem comorbidity (4% of total sample), 20 displayed only behavior problem comorbidity (16% of total sample), and 30 displayed comorbid cognitive and behavior problems (25% of total sample).

Medication status information was available for 81% ($n = 99$) of participants and indicated that 24% ($n = 29$) of participants were taking psychoactive medication (of which 86% [$n = 24$] were stimulants); 57% ($n = 70$) of participants were not taking medication. Medication

Table 1. *Means and Standard Deviations for the GATSB, WISC–III, WJ–R, WRAML, and CBCL Scores (N = 122)*

	M	*SD*	Minimum Score	Maximum Score
GATSB				
Total Problems	60.0	15.4	41	99
Avoidance	58.2	11.9	40	88
Inattentiveness	56.7	13.7	42	99
Uncooperative Mood	57.9	13.8	42	87
WISC–III				
Full-Scale IQ	97.7	15.5	54	135
WJ–R				
Broad Cognitive Ability	98.5	15.4	68	141
WRAML				
General Memory Index	91.4	15.4	52	120
CBCL				
Total Problems	60.8	9.7	36	82
Internalizing	57.6	9.7	34	85
Externalizing	58.5	11.1	32	88

Note: GATSB = Guide to the Assessment of Test Session Behavior; CBCL = Child Behavior Checklist; WISC–III = Wechsler Intelligence Scale for Children–Third Edition; WJ–R = Woodcock–Johnson Psychoeducational Battery–Revised; WRAML = Wide Range Assessment of Memory and Learning; *T* scores ($M = 50$, $SD = 10$) are reported for the GATSB and CBCL; standard scores ($M = 100$, $SD = 15$) are reported for the WISC–III, WJ–R, and WRAML.

Table 2. *Frequency of Cognitive (COG) and Behavioral (BEH) Problems (N = 122)*

		Pure Problems		Comorbid Problems[a]		
	No Problem	COG	BEH	COG	BEH	COG & BEH
Discrete cases[b]	37	15	15	5	20	30
Cognitive problems						
Mental retardation		1		1		5
Math learning disability		3		3		15
Reading learning disability		2		1		6
Writing learning disability		9		5		20
Behavior problems						
Withdrawn behavior			1		5	7
Somatic complaints			2		2	4
Anxious/depressed behavior			2		5	4
Social problems			1		10	13
Thought problems			1		10	8
Attention problems			7		19	21
Delinquent behavior			0		5	8
Aggressive behavior			1		11	12

Note: Mental retardation was defined as IQ ≤ 70 on Wechsler Intelligence Scale for Children–Third Edition (WISC–III) or Woodcock–Johnson Psychoeducational Battery–Revised (WJ–R); learning disability was defined by 15-point ability-achievement discrepancy on WISC–III or WJ–R; behavioral problems were defined by a *T* score ≥ 70 on the Child Behavior Checklist.

[a]The sum of these specific frequencies exceed the discrete case totals because comorbid cases are represented in multiple rows. [b]These frequencies represent discrete cases that sum to the overall sample size ($N = 122$).

was maintained as usual through testing to promote optimal performance, to avoid potential rebound effects from medication cessation, and to promote generalization of findings to the classroom setting. Children taking medication received somewhat higher scores on the GATSB (M_{total} = 65, $M_{avoidance}$ = 64, $M_{inattention}$ = 59, $M_{uncooperative}$ = 60) and the CBCL (M_{total} = 64, $M_{internalizing}$ = 60, $M_{externalizing}$ = 61) than the unmedicated group (GATSB: M_{total} = 58, $M_{avoidance}$ = 56, $M_{inattention}$ = 55, $M_{uncooperative}$ = 56; CBCL: M_{total} = 59, $M_{internalizing}$ = 56, $M_{externalizing}$ = 57). To promote the representativeness of this sample to routine clinic functioning, children were included in overall analyses regardless of their medication status. Validity correlations are also reported separately for the medication and no medication groups.

Materials

CBCL–Parent Report Form (Achenbach, 1991). The CBCL is a 113-item child behavior problem checklist completed by parents. It provides broadband and narrowband scales. The broadband scales measure an Internalizing factor and an Externalizing factor. The narrowband scales measure the following dimensions: Withdrawn Behavior, Somatic Complaints, Anxious/Depressed Behavior, Delinquent Behavior, Aggressive Behavior, Social Problems, Thought Problems, and Attention Problems. Achenbach (1991) reported acceptable internal consistency (α = .90 internalizing, α = .93 externalizing) and test–retest reliability (1-week r = .89, .93; 1-year r = .79, .87; 2-year r = .70, .86) for the CBCL. Achenbach also reviewed numerous studies supporting the validity of the CBCL relative to other parent-report behavior checklists, clinic-referral status, and categorical psychiatric diagnosis. *T* scores were used in all analyses.

Woodcock-Johnson Revised (WJ–R; Woodcock & Johnson, 1989). The WJ–R is a wide-range set of tests for measuring intellectual development, school aptitude, and achievement. The achievement subtests measure accomplishment in the following areas: reading, mathematics, written language, knowledge, and skills. The broad cognitive ability scale measures general intellectual abilities and is the average of abilities across standard subtests. The standard scale of broad cognitive ability has a median (across ages) reliability of .95 with a 2-point standard error of measurement. Concurrent validity, as measured by comparing the WJ–R with several other measures of cognitive ability, was a minimum of .48 in the 3-year-old age group, .57 in the 9-year-old age group, and .64 in the 17-year-old age group (Woodcock & Johnson, 1989). A third edition of the Woodcock–Johnson battery was released following data collection for this study. Standard scores were used in all analyses.

Wechsler Intelligence Scale for Children–Third Edition (WISC–III; Wechsler, 1993). The WISC–III is a standardized test of intellectual ability for children ages 6 to 16 years. The WISC–III is an objective measure of intellectual ability comprised of 13 subtests. The core 11 subtests correlate with full scale IQ (FSIQ) for all ages by a minimum of .56. Reliability for the WISC–III FSIQ was .96 (average for all ages). Additionally, the WISC– III and the GATSB used the same standardization sample of 2,200 children ages 6 to 16. The ethnic proportions of the sample were based on the 1988 census survey. IQ scores were used in all analyses.

Wide Range Assessment of Memory and Learning (WRAML; Adams & Sheslow, 1990). The WRAML is a test of children's memory and learning that consists of nine subtests and yields a general memory index and three subscales: Verbal Memory, Visual Memory, and Learning. The WRAML has a minimum subtest reliability of .78. Standard scores were used in all analyses.

Guide to the Assessment of Test Session Behavior (GATSB; Glutting & Oakland, 1993). The GATSB is a 29-item standardized rating form completed by the clinician immediately following the administration of the WISC–III or WIAT. Clinicians rate the child's behavior on a 3-point Likert scale. The GATSB yields an overall problem score that is composed of three primary factor scores: Avoidance, Inattentiveness, and Uncooperative Mood. Normative data for the GATSB were obtained in conjunction with the WISC–III and WIAT standardization sample. Norms are provided for three age groups: 6 to 9 years, 9 to 13 years, and 13 to 16 years. The GATSB scales have demonstrated adequate internal consistency (α = .84 to .92) and test–retest reliability (r = .71 to .87), consistent factorial structure, and correlations of approximately –.30 with WISC–III and WIAT scores. In a subsample of 50 participants from our clinic, the GATSB displayed acceptable internal consistency with α coefficients of .97 for the total score, and .93, .92, and .92 for the Uncooperative Mood, Avoidance, and Attention subscales, respectively. *T* scores were used in all analyses.

Procedure

Participants in this study were assessed using a comprehensive battery that was administered by master's- or doctoral-level psychology staff (see Newcomb & Drabman, 1995, for details). A generic informed consent was obtained from the parent or legal guardian during the clients' initial visit to the medical center, and specific consent for the assessment procedures was obtained prior to testing. The psychoeducational assessment battery was generally completed during two separate 4-hr sessions that occurred on consecutive days. The WISC–III and GATSB were administered during

the first testing session, along with those portions of the WJ–R that require administration at two points in time (e.g., the encoding portions of long-term retrieval tasks). The remainder of the testing was completed in the following session. Another psychologist, who was not aware of the GATSB scores, administered the CBCL during a clinical interview with the child's parent(s). Thus, different clinicians administered the parent ratings and cognitive measures (i.e., WISC–III and WJ–R). The use of three different methods (parent-report, clinician-rating, and child performance) to measure the three primary constructs of interest (behavior problems, test-session behavior, and cognitive ability) promoted the independence of scores and represented routine clinic procedure.

We examined the validity of the GATSB using several analyses. For the sake of brevity, only a subset is reported. First, because minimal differences were observed between zero-order correlations and partial correlations controlling for age and sex, only the partial correlations are reported. Second, because a generally consistent pattern of findings emerged between the overall ability indexes from cognitive measures and subscales (e.g., verbal and performance IQ from the WISC–III, factor scores from the WJ–R, and so on), only the overall scales are described. Exceptions to this pattern were apparent on the long-term retrieval scale of the WJ–R, and the visual and learning indexes from the WRAML, so these are described further. Similarly, for the CBCL we present broadband but not narrowband internalizing and externalizing scales, as well as findings for the attention problems, social problems, and thought problems scales.

Results

Our first question examined the intrasession validity correlations of the GATSB with children's performance on the WISC–III. We then explored the correlations between the GATSB and scores on other cognitive measures. Exosession validity was examined by calculating correlations between the GATSB and the CBCL. Next, we investigated the incremental validity of each GATSB subscale over the other GATSB subscales, through partial correlations that controlled for the other GATSB subscales in addition to age and sex. Finally, we separately examined the GATSB's validity within the group of children who were receiving medications and the group who was not. An α level of .05 was used in all subsequent analyses.

Concurrent Validity

Consistent with prior studies, all of the GATSB indexes were significantly correlated with the WISC–III FSIQ score (see Table 3). These correlations provided further evidence of the intrasession validity of the GATSB.

A similar but somewhat more variable pattern of correlations was evident for the other cognitive validity measures. Notably, all of the observed correlations between GATSB indexes and the cognitive measures were

Table 3. *Concurrent and Incremental Validity of the GATSB as Indicated by Partial Correlations With the WISC–III, WJ–R, WRAML, and CBCL Validity Indexes After Controlling for Sex and Age, and Sex, Age, and the Other GATSB Scales (N = 122)*

	GATSB Scale			
	Total	**Avoidance**	**Uncooperative Mood**	**Inattention**
WISC–III				
Full Scale IQ	–.46**	–.47**/–.28**	–.39**/–.03	–.28**/–.11
WJ–R				
Broad Cognitive Ability	–.41**	–.40**/–.23*	–.32**/.05	–.39**/–.20*
Broad Reading	–.30**	–.35**/–.29**	–.17/.19*	–.28**/–.18*
Broad Math	–.27**	–.33**/–.27**	–.18/.10	–.21*/–.08
Broad Written Language	–.30**	–.35**/–.28**	–.17/.23*	–.34**/–.27**
WRAML				
General Memory Index	–.24**	–.29**/–.20*	–.20*/.06	–.24*/–.11
CBCL				
Total Problems	.23*	.19*/.03	.25**/.12	.20*/.03
Internalizing	.12	.15/.11	.10/–.02	.10/.03
Externalizing	.19*	.10/–.08	.22*/.19*	.15*/.00
Attention Problems	.32**	.25**/.03	.31**/.09	.33**/.15
Social Problems	.27**	.18/–.05	.30**/.18*	.25**/.06
Thought Problems	.26**	.31**/.22*	.21*/–.02	.20*/.03

Note: GATSB = Guide to the Assessment of Test Session Behavior; WISC–III = Wechsler Intelligence Scale for Children–Third Edition; WJ–R = Woodcock–Johnson Psychoeducational Battery–Revised; WRAML = Wide Range Assessment of Memory and Learning; CBCL = Child Behavior Checklist; Coefficients following the slash (/) character are partial correlations controlling for gender, age, and other GATSB scales.
*$p \leq .05$. **$p \leq .01$.

negative (see Table 3). However, correlations for the achievement scales from the WJ–R (i.e., reading, math, and written language) and the memory index from the WRAML were somewhat lower in magnitude in comparison to the cognitive ability indexes. Further, not all of the correlations were significantly different from zero. On the WJ–R, the Uncooperative Mood subscale of the GATSB did not significantly correlate with any of the achievement scales. Nonsignificant partial correlations were also observed on the WJ–R long-term storage and retrieval scale (pr = –.10, –.11, –.01, –.17 for GATSB Total, Avoidance, Uncooperative Mood, and Inattention, respectively). Relative to the general memory index on the WRAML, partial correlations were larger and significant for the verbal memory index (pr = –.33, –.39, –.27, –.29), but were not significant for the visual memory (pr = –.10, –.16, –.09, –.10) and learning index (pr = –.17, –.18, –.13, –.22), except that GATSB Inattention significantly correlated with the learning index (pr = –.22, p = .017). These results generally support the validity of the GATSB, but we advise caution in the use of GATSB scores for moderating interpretations about memory scales.

Consistent with previous research, we found limited evidence for the exosession validity of the GATSB. The predicted convergent correlations were significant for the GATSB Total Problems with the CBCL Total Problems scale, the GATSB Uncooperative Mood scale with the CBCL Externalizing Problems scale, and the GATSB Inattention scale with the CBCL Attention Problems scale. The GATSB Avoidance subscale was not significantly correlated with the CBCL Internalizing subscale. Discriminant exosession validity for the GATSB was also limited. A variety of additional significant correlations were observed between the GATSB scales and the CBCL Attention, Social, and Thought Problems scales. Thus, the observed pattern of results suggested that the GATSB test session behavior scores were related in a nonspecific fashion to parent-report behavior problems outside of the testing setting but were not significantly related to children's internalizing problems.

Incremental Validity

To examine the incremental validity of each individual GATSB subscale, partial correlations were calculated for each subscale after removing the effects due to the other two subscales, age and sex. For example, the incremental validity of the GATSB Avoidance subscale relative to the WISC–III FSIQ was examined by calculating a partial correlation between these scales while controlling for age, sex, GATSB Uncooperative Mood, and GATSB Inattention. Results indicated that GATSB Avoidance made a unique contribution in the prediction of all of the cognitive indexes (see Table 3). GATSB Inattention made a unique contribution for the WJ–R scores, but not for the WISC–III or WRAML scores. Finally, significant partial correlations were observed for GATSB Uncooperative Mood with WJ–R Reading and Written Language, but in the opposite direction of all other scales. Thus, after controlling for the significant contributions of both the Avoidance and Inattention scales, children displaying a more uncooperative mood scored higher on the WJ–R achievement tests.

Consistent with the concurrent validity analyses, examination of the incremental exosession validity of the GATSB provided evidence that the GATSB scales made a nonspecific contribution to the prediction of child behavior problems. The exceptions to this were that the GATSB Uncooperative Mood scale made a unique contribution to the prediction of Externalizing Problems and Social Problems, and the GATSB Avoidance scale made a unique contribution to the prediction of Thought Problems. No significant partial correlations were observed between the GATSB Inattention scale and the CBCL scales.

Medication Effects

Because of the important role that medication may play in altering children's test session behavior, our final set of analyses examined the concurrent validity of the GATSB scales for the group of children receiving medication (n = 29) and the group of children who were not (n = 70). The moderational effect of medication status was tested using a general linear model analysis with one categorical independent variable (medication group), one continuous independent variable (GATSB scores), and two covariates (sex and age). In this analysis, the moderational role of medication status was represented in the medication group by GATSB interaction term. Partial R^2 values were reported as effect size estimates. To maintain consistency with prior analyses, we report partial correlations controlling for age and sex and hypothesis tests of whether these partial correlations were significantly different from zero within each medication group (see Table 4). Due to the differences in sample sizes across groups, comparisons between groups should be made cautiously, because the power of hypothesis testing is highly sensitive to sample size, whereas effect size estimates are less affected by sample size variations (e.g., Keppel, 1991). Findings were similar for the GATSB subscales, so only results from the GATSB total scores are reported.

Medication status was only a significant moderator of the GATSB–cognitive ability relation for the WJ–R Broad Reading scale (pR^2 = .055, p = .025). The partial correlation of the GATSB with the WJ–R Broad Reading scale was significantly greater within the medica-

Table 4. *Concurrent Validity of the GATSB Total Scores for Children Taking (n = 29) and Not Taking Medications (n = 70) as Indicated by Partial Correlations of the GATSB Scales with the WISC–III, WJ–R, WRAML, and CBCL Validity Indexes After Controlling for Sex and Age*

	Moderation Effect Size	Child Groups	
	pR^2	Medication (n = 29)	No Medication (n = 70)
WISC–III			
Full Scale IQ	.002	–.46*	–.38**
WJ–R			
Broad Cognitive Ability	.005	–.54**	–.39**
Broad Reading	.055*	–.63**	–.17
Broad Math	.015	–.38*	–.10
Broad Written Language	.005	–.37	–.24*
WRAML			
General Memory Index	.012	–.32	–.14
CBCL			
Total Problems	.000	.19	.23
Internalizing	.000	.13	.12
Externalizing	.001	.22	.19
Attention Problems	.009	.15	.38**
Social Problems	.000	.26	.25*
Thought Problems	.001	.13	.24*

Note: pR^2 = partial R square for medication by GATSB interaction effect; GATSB = Guide to the Assessment of Test Session Behavior; WISC–III = Wechsler Intelligence Scale for Children–Third Edition; WJ–R = Woodcock–Johnson Psychoeducational Battery–Revised; WRAML = Wide Range Assessment of Memory and Learning; CBCL = Child Behavior Checklist.
*$p \leq .05$. **$p \leq .01$.

tion group (–.63) than within the no-medication group (–.17). Further examination of the partial correlations revealed that within both the medication and no-medication groups, the intrasession validity of the GATSB was evidenced by significant correlations between the GATSB total score and the WISC–III that were of comparable magnitude. Similarly, the GATSB was significantly correlated with the WJ–R Broad Cognitive Ability scale within both groups. However, for the other cognitive measures, the partial correlations of the GATSB with the WJ–R and WRAML were in the –.32 to –.38 magnitude range for the medication group, but ranged from –.24 to –.30 in the total sample.

Medication status did not significantly moderate any exosession validity relations. Further examination of the partial correlations revealed that the GATSB performed quite consistently across the medication and no-medication groups relative to the CBCL Total, Internalizing, Externalizing, and Social Problem scales. The largest variation was on the CBCL Attention Problems scale, but this was still a small, nonsignificant effect. In the no-medication group, the partial correlation between the GATSB and the CBCL Attention Problems (.38) was significantly different from zero, but in the medication group this correlation was smaller in magnitude (.15) and not significant. A similar but even smaller effect was observed on the Thought Problems scale. Altogether, the GATSB performed relatively consistently across medication and no-medication groups of children with respect to its concurrent validity, except for the measure of reading achievement.

Discussion

Consistent with prior studies, the GATSB demonstrated adequate intrasession validity relative to the WISC–III. Moreover, correlations between the GATSB and other cognitive measures (i.e., WJ–R and WRAML) were of comparable magnitude to the intrasession correlation with the WISC–III. These results provided converging evidence for the validity of using GATSB scores obtained from the WISC–III to moderate interpretations about children's performance on the WISC–III and other cognitive measures (with the exception of memory scales). The pattern of GATSB subscale correlations provided circumscribed support for the practice of drawing unique inferences from the Avoidance, Uncooperative Mood, and Inattention scales. The GATSB Avoidance scale played the most prominent role and accounted for significant unique variance in the cognitive measures, and the GATSB Inattention scale made an additional contribution to some of the WJ–R scales. Altogether, these findings support the use of the GATSB total score as a global indicator of test behavior problems that are associated with lower scores across cognitive testing. They also highlight avoidance as the most prominent and uncooperative mood as the least prominent test behavior factors.

Consistent with earlier studies, the intrasession validity of the GATSB was estimated in approximately the –.30 to –.40 range (Glutting & Oakland, 1993; Konald et al., 1998; Maller et al., 1998). If these correlations are considered in the context of a standardized

regression equation, they suggest that for every 10-point increment in GATSB *T* scores, clinicians should expect a .3 to .4 standard deviation reduction in cognitive test scores (e.g., a 4.5- to 6-point reduction in standard scores with a $SD = 15$). In some cases (e.g., IQs around average), this range of fluctuation may not be of clinical significance, and clinicians can acknowledge in reporting that observed scores may be underestimates of true scores and perhaps extend the range of their reported confidence intervals. However, when cases fall into borderline ranges or when making decisions with high potential impact on the child's life (e.g., special education placement, diagnoses of mental retardation or learning disability, and so on), even relatively mild GATSB elevations may lead to adjusted ability estimates that traverse relevant cutting scores. Therefore, clinicians may wish to suspend judgment and cite GATSB scores as evidence for potentially elevated levels of testing error. In this regard, use of the GATSB may help estimate a range over which to moderate interpretation and prevent completely discarding test scores based on any evidence of problematic test behavior.

These findings provide limited evidence for the exosession validity of the GATSB. The GATSB total score significantly correlated with the parent-reported CBCL total score at a level (.23) that was somewhat below the typical level of agreement between multiple reporters (e.g., Achenbach, 1991). Few of the convergent and discriminant correlations emerged as predicted for the broadband and narrowband scales. Instead, a generalized pattern of modest correlations was observed between the GATSB and CBCL scales, with the exception of the Internalizing scales. The GATSB subscales accounted for little unique variance in behavioral indexes, and only the Uncooperative Mood scale displayed the predicted convergent relation. At present, these findings advise against drawing inferences about a child's behavior outside of the test setting from the child's behavior during testing. Although such low magnitude correlations might be expected based on the novelty of the testing environment, they indicated that the GATSB might have limited practical significance for exosession use. In a clinical setting, such inferences are best based on observations of the child's behavior in various settings and on information obtained from multiple reporters of the child's behavior in those settings.

Four issues are important to consider when interpreting these findings. First, the GATSB may be influenced by a variety of child difficulties and therefore fail to reflect specificity in the type of underlying emotional or behavioral problem. Second, these conclusions should be viewed tentatively because this sample represented clinic-referred children, who tend to display elevated levels of comorbid psychopathology (Kazdin & Kagan, 1994). Forty-five percent of the sample was characterized by comorbid cognitive problems, behavioral problems, or both. Such comorbidity would tend to provide overly stringent tests of discriminant validity and tend to inflate subscale intercorrelations. This study used a dimensional taxonomy of emotional and behavioral problems (Achenbach, 1991) and did not examine psychiatric diagnoses. Investigation of diagnostic categories may yield a different pattern of results on the GATSB. Third, because clinicians completed the GATSB following administration of the WISC–III, GATSB scores may be affected by the clinician's impression of the child's overall level of intelligence even though the child's actual IQ score was not yet calculated. In other words, not only may poor behavior interfere with task performance, but also children who perform well on cognitive tests may be rated as better behaved. Nevertheless, completion of the GATSB following observations during an intelligence test is consistent with its intended clinical use and prior validation studies.

Fourth, this study examined the validity of the GATSB under routine clinic procedures. Several methodological limitations of the study should be addressed in future research. For example, the WISC–III and GATSB were administered prior to the other cognitive measures. This strategy does not allow for examination of order effects. Because the same clinician administered the GATSB and the subsequent cognitive measures, correlations between these measures may tend to be inflated. Future studies could counterbalance the order of administration of the GATSB and validity criterion to control this potential bias. This study examined a very heterogeneous clinical sample. Future studies may examine a host of potential moderating variables such as age, intelligence, and reason for referral.

Another logical next step in the validation of the GATSB is to examine GATSB scores based on observations of child behavior during tasks other than those on the WISC–III or WIAT. The weakest validity correlations observed in this study were on the memory indexes. Part of the explanation for the GATSB's weakness in this domain may be that the Digit Span subtest is the only primary memory task of the WISC–III, and, therefore, clinicians do not have an adequate sample of the child's behavior during encoding and recall tasks. This speculation highlights the merits of future research investigating the use of the GATSB with observations made during administration of other psychoeducational batteries, particularly those that sample a broader domain than the WISC–III. Such investigations will help establish the generality and stability of children's behavior throughout administration of a test battery as well as expand the utility of the GATSB to assessments that do not include administration of the WISC–III or WIAT scales.

Finally, future discussion is warranted regarding the interpretive use of structured behavioral observations

such as the GATSB. Considerable evidence suggests that much child behavior is context specific (e.g., Achenbach, McConaughy, & Howell, 1987). The present findings suggested that sufficient contextual similarity existed across cognitive test batteries to support predicting subsequent cognitive task performance from behavior during a preliminary battery (i.e., WISC–III). However, this study did not address the extent to which child behavior fluctuates significantly across subtests. Insufficient evidence exists to evaluate the use of the GATSB as a measure of behavior during individual subtests, for example, as an index of fatigue or level of item difficulty.

In sum, this study extended the validation of the GATSB to a clinical sample using cognitive and achievement criterion beyond the WISC–III and WIAT measures. The significant intrasession validity coefficients of the GATSB support its use as a broad index of behavioral challenges during individual testing. Current evidence suggests that the GATSB should be used as a measure of total test session performance and that subscales should be given less weight and interpreted cautiously. Further, the GATSB displayed only modest relations with parent-reported behavior problems, and therefore it is not recommended as an indicator of behavior outside of the test setting.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 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.

Adams, W., & Sheslow, D. (1990). *Wide Range Assessment of Memory and Learning.* San Antonio, TX: Psychological Corporation.

Glutting, 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). *Guide to the Assessment of Test Session Behavior: Manual.* San Antonio, TX: Psychological Corporation.

Glutting, J., Oakland, T., & Konald, T. (1994). Criterion-related bias with the *Guide to the Assessment of Test-Session Behavior* for the WISC–III and WIAT: Possible race/ethnicity, gender, and SES effects. *Journal of School Psychology, 32,* 355–369.

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

Glutting, J., Robins, P., & de Lancey, E. (1997). Discriminant validity of test observations for children with attention deficit/hyperactivity. *Journal of School Psychology, 35,* 391–401.

Kazdin, A. E., & Kagan, J. (1994). Models of dysfunction in developmental psychopathology. *Clinical Psychology: Science and Practice, 1,* 35–52.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Konald, T., Glutting, J., Oakland, T., & O'Donnell, L. (1995). Congruence of test-behavior dimensions among child groups that vary in gender, race/ethnicity, and SES. *Journal of Psychoeducational Assessment, 13,* 111–119.

Konald, T. R., Maller, S. J., & Glutting, J. J. (1998). Measurement and non-measurement influences of test-session behavior on individually administered measures of intelligence. *Journal of School Psychology, 36,* 117–132.

Maller, S. J., Konald, T. R., & Glutting, J. J. (1998). WISC–III factor invariance across samples of children exhibiting appropriate and inappropriate test-session behaviors. *Educational and Psychological Measurement, 58,* 467–474.

Newcomb, K. P., & Drabman, R. S. (1995). Child behavioral assessment in the psychiatric setting. In R. T. Ammerman & M. Hersen (Eds.), *Handbook of child behavior therapy in the psychiatric setting* (pp. 3–25). New York: Wiley.

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

Wechsler, D. (1993). *Wechsler Intelligence Scale for Children–Third Edition.* San Antonio, TX: Psychological Corporation.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psycho-Educational Battery–Revised.* Itasca, IL: Riverside.

Received June 11, 2001
Accepted December 7, 2001