Psychological Assessment
2004, Vol. 16, No. 2, 133–138

1040-3590/04/$12.00 DOI: 10.1037/1040-3590.16.2.133

Temporal Stability of WISC–III Subtest Composite: Strengths and Weaknesses

Marley W. Watkins
Pennsylvania State University, University Park Campus

Gary L. Canivez
Eastern Illinois University

The Wechsler Intelligence Scale for Children—Third Edition (D. Wechsler, 1991; WISC–III) is often used to identify subtest-based cognitive strengths and weaknesses that are subsequently used to generate interventions. Given that intelligence is presumed to be an enduring trait, cognitive strengths and weaknesses identified via subtest analysis should also be stable over time. This was evaluated with 579 students who were twice tested with the WISC–III. Based on 66 subtest composites, 6 or 7 interpretable cognitive strengths and weaknesses were found on each WISC–III administration. However, subtest-based strengths and weaknesses replicated across test–retest occasions at chance levels ($Mdn_{\kappa} = .02$). Because subtest-based cognitive strengths and weaknesses are unreliable, recommendations based on them will also be unreliable.

The Wechsler scales are the most popular individual measures of intelligence for children, adolescents, and adults (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Belter & Piotrowski, 2001). Among the school-age population, millions of children have been administered the Wechsler Intelligence Scale for Children—Third Edition (WISC–III; Wechsler, 1991) as part of an evaluation to determine eligibility for special education services (Kamphaus, Petoskey, & Rowe, 2000). Beyond its diagnostic applications, the WISC–III is often used to identify cognitive strengths and weaknesses that form the basis for psychoeducational recommendations (Zeidner, 2001).

On the basis of these principles, intricate subtest interpretation systems (Kaufman, 1994; Sattler, 2001) have achieved wide popularity in psychology training and practice (Alfonso et al., 2000; Groth-Marnat, 1997; Kaufman, 1994; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). These interpretative strategies typically begin at the top of the WISC–III score hierarchy by making inferences about general abilities from the Full-Scale IQ (FSIQ), Verbal IQ (VIQ), and Performance IQ (PIQ) scores only if subtest scores do not vary significantly. If there is substantial scatter among lower level components, then an IQ score "represents a summary of diverse abilities and does not represent a unitary entity" (Kaufman & Lichtenberger, 2000, p. 424).

Following this logic, specific patterns of subtest scores are presumed to substantially invalidate global intelligence indices (Groth-Marnat, 1997) so that subtest scores and subtest composites, rather than IQ composites, become the focus of interpretation. Subtests that are significantly higher or lower than the child's own average (ipsative comparisons) are deemed relative strengths or weaknesses, respectively. Next, hypotheses concerning the underlying causes of significant subtest variations are identified by locating abilities thought to be shared by two or more subtests. Extensive lists of the abilities presumed to underlie each subtest are provided by Kaufman (1994), Sattler (2001), and Kaufman and Lichtenberger (2000). Finally, these hypotheses are used to generate educational and psychological interventions and remedial recommendations.

Intelligence is presumed to be an enduring trait (Schuerger & Witt, 1989), and it is expected that scores on tests measuring intelligence should produce high stability coefficients and nonsignificant mean differences when compared across a long time interval ($>$ 1 year). In a series of studies, Canivez and Watkins (1998, 1999, 2001) demonstrated the substantial long-term (3-year) stability of WISC–III IQ scores and the Verbal Comprehension and Perceptual Organization factor index scores. The stability of these indexes were invariant across several demographic variables (age, race, and gender) as well as across various disability groups.

An individual's cognitive strengths and weaknesses identified via ipsative subtest analysis should also be stable over time if such patterns or characteristics are to have clinical utility. As noted by Cronbach and Snow (1977), "any long-term recommendations as to a strategy for teaching a student would need to be based on aptitudes that are likely to remain stable for months, if not years" (p. 161). It is clear that popular subtest interpretation systems render recommendations that are long term in nature. They include, for example, suggestions regarding differential teaching styles, curricular materials, and learning environments (Kaufman, 1994; Kaufman & Lichtenberger, 2000).

Marley W. Watkins, Department of Educational and School Psychology and Special Education, Pennsylvania State University, University Park Campus; Gary L. Canivez, Department of Psychology, Eastern Illinois University.

This research was supported, in part, by an Eastern Illinois University faculty development grant and a Pennsylvania State University College of Education Alumni Society faculty research initiation grant. We thank the school psychologists who generously responded to our request for WISC–III data as well as Tim Runge, Lisa Samuels, and Daniel Heupel for assistance in data entry.

Correspondence concerning this article should be addressed to Marley W. Watkins, Pennsylvania State University, University Park Campus, 227 CEDAR Building, University Park, PA 16802. E-mail: mww10@psu.edu

Standards for educational and psychological testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) require that test interpretation methods demonstrate empirical support. Stability of cognitive strengths and weaknesses over time is one type of evidence that could support ipsative subtest interpretation. This temporal stability hypothesis was examined with the Wechsler Intelligence Scale for Children—Revised (WISC–R; Wechsler, 1974) with 303 randomly selected children tested twice as part of WISC–R validation studies and with 189 children twice tested for special education eligibility (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). Classificatory stability of relative cognitive strengths and weaknesses identified by subtest elevations and depressions was near chance levels for both groups of children. Likewise, the multivariate stability of WISC–R subtest profiles across a 3-year test–retest interval for 60 students was too low for clinical use (Livingston, Jennings, Reynolds, & Gray, 2003). These analyses have yet to be replicated with other participants or tests. Consequently, the present longitudinal study was conducted to investigate the temporal stability of WISC–III subtest and composite score strengths and weaknesses among students twice tested in special education evaluations.

Method

Participants

Participants were 579 students (67.2% male, 32.8% female) twice tested with the WISC–III. Race–Ethnicity was 76.3% Caucasian, 5.7% Hispanic–Latino, 14.7% Black–African American, 0.7% Native American–American Indian, and 2.6% Other–missing. Students were determined to be disabled (or not disabled) by multidisciplinary evaluation teams according to state and federal guidelines governing special education classification. Special education diagnosis on initial evaluation included 63.2% learning disability, 7.3% emotional disability, 9.9% mental retardation, 2.6% speech and language disability, 7.8% unspecified, 3.1% not disabled, and 6.1% other disabilities.

The average test–retest interval was 2.80 years ($SD = 0.55$) with a range of 0.50 to 6.00 years. Only 1.1% of the reevaluations occurred less than 1 year following the first evaluation. The mean age of students at first testing was 9.15 years and ranged from 6.00 to 14.60 years. The mean age of students at second testing was 11.96 years and ranged from 7.50 to 16.90 years. Additional detailed demographic information may be obtained from Canivez and Watkins (1998, 1999, 2001).

Instrument

The WISC–III is an individually administered test of intelligence for children aged 6 years through 16 years 11 months. The WISC–III was standardized on a nationally representative sample ($N = 2{,}200$) closely approximating the 1988 United States Census on gender, parent education (socioeconomic status), race–ethnicity, and geographic region. Ten subtests are mandatory for computation of summary IQ scores. Two optional subtests can be administered if factor index scores are to be computed. Extensive evidence of reliability and validity is presented in the WISC–III manual (Wechsler, 1991).

Procedure

Two thousand school psychologists were randomly selected from the National Association of School Psychologists' membership roster and invited to participate by providing test scores and demographic data obtained from recent special education triennial reevaluations. Data were voluntarily submitted on 667 cases by 145 school psychologists from 33 states. Of these cases, 579 contained scores for all 10 mandatory WISC–III subtests.

Specific WISC–III subtest composites were extracted from the literature (Bracken, McCallum, & Crain, 1993; Groth-Marnat, 1997; Kaufman, 1994; Macmann & Barnett, 1997; Siegel & Piotrowski, 1994). The ipsative methods detailed by Kaufman and Lichtenberger (2000) were precisely followed to identify WISC–III subtest ability patterns. Altogether, 12 individual and 54 composite subtest-based abilities were distinguished (see Tables 1 and 2). Following all of Kaufman and Lichtenberger's (2000) decision rules (summarized in their Appendix C), each ability was then categorized as a relative weakness if it was significantly below the appropriate mean subtest scaled score, a relative strength if it was significantly above the mean subtest scaled score, and average if it did not significantly deviate from the mean subtest score. Classificatory stability of these categorizations was then quantified with coefficient kappa (Cohen, 1960), which is a chance-corrected metric for estimating agreement for nominal scale data.

Individual subtest deviations and intermediate classifications were also retained and tested for stability. Following Kaufman and Lichtenberger (2000), several levels of VIQ–PIQ differences, factor discrepancies, and subtest variability (scatter) among IQ composites were also tested for temporal stability with coefficient kappa (see Table 2). The stability of these subtests and intermediate classifications was also evaluated for two subsets of cases: (a) 66 cases with shorter test–retest intervals ($\leq$ 2 years) and (b) 513 cases with longer test–retest intervals ($>$ 2 years).

Results

Descriptive statistics for the WISC–III subtest and composite IQ scores across test and retest occasions are presented in Table 3. Although somewhat lower than the WISC–III standardization sample, IQ scores are consistent with other samples of students with disabilities (Kavale & Nye, 1985–1986). The largest absolute mean difference between test–retest IQ scores was only 2.4 points, and none were significantly different. The largest absolute mean difference between individual subtests was a practically insignificant 0.7 points.

Of the 54 subtest composites analyzed (Table 1), the median kappa coefficient was –.012, with a range of –.119 (Acquired Knowledge) to +.139 (Concentration). The median kappa coefficient for the 12 individual subtests (Table 2) was +.020, with a range of –.031 to +.086. Kappa coefficients for the 10 intermediate categorizations ranged from –.126 to +.099, with a median of +.006 (Table 2).

Interpretative guidelines for kappa have been provided by Cicchetti (1994). Values of less than +.40 represent poor agreement. Of the 76 statistical tests, only 7 were significant at $p < .05$. However, one was for a negative kappa (indicating disagreement), and four coefficients would have been expected to be significant by chance. Thus, it appears that kappa coefficients were generally not different than zero (were at chance levels). Similar poor agreement across time was found when only weaknesses ($Mdn_{\kappa} = .016$) or strengths ($Mdn_{\kappa} = .006$) were considered.

For the 145 cases where all 12 WISC–III subtests were administered on both occasions, 68.3% of the students displayed at least one significant cognitive weakness, and 62.8% demonstrated at least one significant cognitive strength on the first test. On average, these students exhibited 3.8 relative weaknesses and 3.5 relative strengths on their first WISC–III and 4.5 relative weak-

Table 1
Agreement on WISC–III Composite Subtest-Based Strengths and Weaknesses Across a Test–Retest Interval of 2.8 Years

Subtests[a]	Ability[b]	κ
IN, CM	Culture-loaded knowledge	.026
IN, PC	Alertness to environment	−.030
CM, PA	Common sense, social comprehension, social judgment	−.011
AR, PC	Concentration	−.011
IN, VO	Fund of information, foreign language background, intellectual curiosity, richness of environment	−.012
SM, VO	Handling abstract verbal concepts, verbal concept formation, degree of abstract thinking	−.018
VO, CD	Learning ability	−.024
AR, DS	Mental alertness, attention span, hearing difficulties	.031
CD, SS	Motivation level, obsessive concern with accuracy and detail, processing speed	−.005
PA, OA	Nonverbal reasoning	−.018
PA, SS	Planning ability	.079
CD, BD	Reproduction of a model, visual perception of abstract stimuli	−.008
BD, SS	Spatial visualization	−.064
BD, OA	Trial-and-error learning	.057
PC, OA	Verbal directions, holistic (right-brain) processing	−.056
SM, CM	Verbal reasoning, overly concrete thinking	.014
CD, PA	Visual sequencing, convergent production	.080*
PC, PA	Visual perception of complete meaningful stimuli, visual organization	.065
PC, CD	Visual memory	−.034
AR, DS, SS	Attention span	−.036
SM, VO, BD	Concept formation	.044
CD, PA, SS	Convergent production	.105
SM, PC, PA	Distinguishing essential from nonessential detail	−.024
SM, DS, OA	Flexibility	.082*
IN, SM, VO	Interests, extent of outside reading	−.018
VO, CD, SS	Learning ability	.088
IN, AR, VO	Acquired knowledge, long-term memory, school learning	−.119*
CD, OA, SS	Persistence	.024
SM, AR, VO	Semantic cognition	−.017
PC, BD, OA	Simultaneous processing, visual processing spatial, cognitive style (field dependence–independence)	−.013
AR, DS, CD	Sequential processing, symbolic content, facility with numbers, freedom from distractibility, anxiety, distractibility	.041
DS, CD, SS	Short-term memory (auditory or visual)	.091
PA, BD, OA	Synthesis	.032
IN, AR, DS	Memory, little expression required, simple verbal response	−.014
IN, AR, CM	Understanding long questions (or stimuli)	.006
SM, VO, DS	Understanding words (or brief stimuli)	−.013
PC, CD, SS	Visual memory	−.062
PC, PA, OA	Visual perception of meaningful stimuli	−.060
CD, BD, SS	Visual perception of abstract stimuli	−.074
SM, VO, CM	Verbal conceptualization, much expression required, verbal expression	−.028
AR, PC, CD, SS	Concentration	.139*
IN, VO, CM, PA	Cultural opportunities at home	.066*
CD, PA, BD, SS	Complex verbal directions, integrated brain functioning	.081
SM, PC, PA, SS	Distinguish essential from nonessential detail	−.055
AR, DS, CD, SS	Encode information for processing, anxiety, distractibility, learning disabilities–ADHD	.039
PC, BD, OA, SS	Figural evaluation	−.043
SM, CM, DS, PC	Negativism	−.033
AR, DS, PA, CD	Sequencing	.037
CD, BD, OA, SS	Visual–motor coordination, visual–perceptual problems	−.034
IN, SM, VO, CM	Verbal comprehension	−.015
AR, DS, PC, CD, SS	Attention–concentration	.067
IN, SM, VO, CM, PA	Crystallized ability	.045
SM, AR, PA, BD, OA	Fluid ability	−.038
SM, AR, CM, PA, OA	Reasoning	.002
Mdn		−.012

Note. WISC–III = Wechsler Intelligence Scale for Children—Third Edition; IN = Information; CM = Comprehension; PC = Picture Completion; PA = Picture Arrangement; AR = Arithmetic; VO = Vocabulary; SM = Similarities; CD = Coding; DS = Digital Span; SS = Symbol Search; OA = Object Assembly; BD = Block Design; ADHD = attention-deficit/hyperactivity disorder.
[a] $N = 145$ for SS, $N = 418$ for DS, and $N = 579$ for other subtests. [b] From Bracken, McCallum, & Crain, 1993; Kaufman, 1994; Kaufman & Lichtenberger, 2000; Macmann & Barnett, 1997; Siegel & Piotrowski, 1994.
* $p < .05$.

Table 2
Kappa Coefficients for WISC–III Subtests and IQ Components Across Shorter (≤2.0 Years), Longer (>2.0 Years), and Total Test–Retest Intervals

Subtest–IQ component	Shorter interval[a]	Longer interval[b]	Total interval[c]
PC	.062	.039	.041
IN	.163	−.007	.017
CD	.179*	.064	.086*
SM	.049	.039	.041
PA	.050	.088*	.085*
AR	—	−.046	−.026
BD	−.066	−.025	−.031
VO	−.086	.072*	.052
OA	−.041	.012	.006
CM	.014	−.008	−.004
SS	—	−.013	.023
DS	—	.022	.006
VIQ–PIQ (±11 points)	.034	−.023	−.013
VIQ–PIQ (±19 points)	.000	.023	.020
VC–FD (±13 points)	−.016	.039	.030
PO–PS (±15 points)	—	−.097	−.126
VC–PO (±12 points)	.121	−.051	−.024
VC–PO (±19 points)	.340*	.024	.075
VIQ scatter (≥7 points)	−.040	−.002	−.009
PIQ scatter (≥9 points)	.023	.107*	.099*
VC scatter (≥7 points)	−.078	−.003	−.015
PO scatter (≥8 points)	−.051	.091*	.078
Mdn κ	.019	.017	.019
M test–retest interval	1.7	3.0	2.8

Note. WISC–III = Wechsler Intelligence Scale for Children—Third Edition. PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension Index; FD = Freedom From Distractability Index; PO = Perceptual Organization Index; PS = Perceptual Speed Index; scatter = high–low score of each index–IQ; — = sample size too small to calculate.
[a] $N = 13$ for PS; 41 for FD; and 66 for VC, PO, VIQ, and PIQ comparisons. [b] $N = 132$ for PS; 377 for FD; and 513 for VC, PO, VIQ, and PIQ comparisons. [c] $N = 145$ for PS; 418 for FD; and 579 for VC, PO, VIQ, and PIQ comparisons.
* $p < .05$.

nesses and 3.4 relative strengths on retesting. Similarly, an average of 3.3 relative cognitive weaknesses and 2.8 relative cognitive strengths were found for the 579 students who were administered the 10 mandatory subtests on both test and retest occasions.

There were insufficient participants to reliably analyze subgroups of cases for the 54 composite subtest-based strengths and weaknesses identified in Table 1, but the 12 individual subtests and 10 intermediate IQ components were separately analyzed for cases with shorter (≤ 2 years) and longer (> 2 years) test–retest intervals (see Table 2). Agreement was poor for both intervals (Mdn_{κ} = .019 and .017). Eight kappa coefficients decreased in magnitude across time, and nine increased. Thus, the length of the test–retest interval did not appear to impact the results.

The stability of other constructs among these students across the total test–retest period was also examined. For example, classificatory agreement of students into exceptional child categories (i.e., learning disability, emotional disability, and mental retardation) across time was excellent (*k*s = .75, .75, and .91, respectively). Dichotomizing FSIQ at a cut score of 70 produced fair to good agreement ($k = .52$) on test and retest. Agreement was also fair to good when reading and mathematics scores were dichotomized using cut scores of 85 (*k*s = .501 and .508, respectively).

Discussion

Psychologists often proffer interventions and remedial recommendations based on hypotheses about WISC–III subtest and subtest composite scores (Kaufman, 1994; Pfeiffer et al., 2000). The present study investigated the temporal stability of WISC–III subtest and subtest composite scores among students twice tested for special education eligibility. On average, subtest-based cognitive strengths and weaknesses replicated across test–retest occasions at chance levels. None of the 66 subtest composites reached the minimal level of agreement necessary for clinical use (Cicchetti, 1994). Stability was not improved when cases with a shorter test–retest interval were analyzed. As was found with the WISC–R (McDermott et al., 1992), the long-term stability of WISC–III subtest composites among students with disabilities was poor.

Further, the large number of possible subtest recategorizations ensures that several significant ability strengths or weaknesses will be identified for most children. In the present study, an average of six or seven interpretable WISC–III subtest and subtest composite scores was found for each student. Thus, it is highly likely that an

Table 3
Means and Standard Deviations of WISC–III Subtest and IQ Scores Across a Test–Retest Interval of 2.8 Years

Subtest–IQ Component	Test		Retest		
	M	*SD*	*M*	*SD*	Difference
VIQ	88.80	15.93	88.20	15.77	−0.6
PIQ	90.90	16.71	90.70	17.73	−0.2
FSIQ	88.80	16.15	88.40	16.89	−0.4
VC	90.40	15.78	89.90	15.73	−0.5
PO	91.80	17.07	92.70	18.36	+0.9
FD	85.60	14.62	85.40	13.46	−0.2
PS	93.10	16.74	90.70	14.85	−2.4
PC	8.70	3.34	9.10	3.35	+0.4*
IN	7.80	3.17	8.00	3.14	+0.2
CD	8.30	3.43	7.60	3.21	−0.7*
SM	8.20	3.41	8.40	3.27	+0.2
PA	8.50	3.58	8.70	3.93	+0.2
AR	7.30	3.14	7.20	2.92	−0.1
BD	8.40	3.72	8.30	4.01	−0.1
VO	8.00	3.23	7.50	3.12	−0.5*
OA	8.40	3.38	8.50	3.58	+0.1
CM	8.70	3.70	8.40	3.55	−0.3
SS	8.60	3.79	8.60	3.47	0.0
DS	7.30	2.72	7.40	2.70	+0.1

Note. $N = 145$ for SS and PS, 418 for DS and FD, and 579 for other subtests and IQs. WISC–III = Wechsler Intelligence Scale for Children — Third Edition. VIQ = Verbal IQ; PIQ = Performance IQ; FSIQ = Full-Scale IQ; VC = Verbal Comprehension Index; PO = Perceptual Organization Index; FD = Freedom From Distractability Index; PS = Perceptual Speed Index; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.
* $p < .05$.

examiner will find cognitive strengths and weaknesses to interpret in a WISC–III profile.

Kaufman (1994) opined that a cognitive pattern that is supported by clinical observations or other information "becomes reliable by virtue of its cross-validation" (p. 31). However, suggestions that unreliable cognitive subtest scores somehow become valid when integrated informally and subjectively with a complex mixture of other assessment data are contradicted by the research literature (Dawes, Faust, & Meehl, 1989). To the contrary, it is well known that psychologists are most vulnerable to decision errors in exactly this situation (Davidow & Levinson, 1993; Faust, 1986, 1990). As described by Faust (1990), this "common belief in the capacity to perform complex configural analysis and data integration might thus be appropriately described as a shared professional myth" (p. 478).

As with all research, results of the present study must be considered within the limitations of its design and sample. First, generalization of results may be limited because these WISC–III data were not obtained by random selection. However, the large number of WISC–III cases from across the United States should, to some extent, reduce this threat because it is unlikely that any one type of student would be systematically or preferentially selected. Second, there was no way to validate the accuracy of WISC–III test scores. Although internal consistency of composite scores was checked during data entry, administration, scoring, or reporting errors could have influenced results. Third, the use of reevaluation cases means that those students who were no longer enrolled in special education were not reevaluated and thus not part of the sample.

A final limitation to this study is its duration. A 3-year test–retest interval means that "real changes could take place in cognitive abilities, and you can't be sure what portion of the variance is due to instability of the subtests versus reliable changes in ability" (Kaufman, 1994, p. 32). Although not directly testable, evidence from the cases with shorter test–retest intervals suggested that length of test–retest interval was not a major influence (Table 2). Additionally, Table 3 reveals that IQ scores were stable across time. If cognitive weaknesses had been remediated in special education classes, then IQ scores should have risen to reflect improved performance on previously deficient subtests.

In contrast, the stability of other constructs among these students was fair to excellent. For example, classificatory agreement of students into exceptional child categories across time was excellent. Even the stability of decisions based on IQ and achievement test cut scores was fair to good. Thus, there was considerable evidence that general cognitive and academic skills were reasonably stable across time. Temporal instability was restricted to ipsative analyses. When taken as a whole, this evidence indicates there was little reliable change in ability and suggests that "most of the differential profile patterns are really little more than unrecognized error variance" (Thorndike, 1994, p. 178).

The present results have unambiguous implications for psychological practice. First, because ipsative subtest categorizations are unreliable, recommendations based on them will also be unreliable. Procedures that lack reliability cannot be valid. Second, because most students will exhibit several relative cognitive strengths and weaknesses, the presence of subtest patterns should not be interpreted as unusual or pathognomonic. Third, by beginning the clinical decision-making process with an essentially random component (i.e., the subtest profile) and then searching for confirmation, the psychologist cannot increase, and will likely decrease, judgment accuracy when trying to detect a low-prevalence strength or weakness (Meehl & Rosen, 1955). Thus, flawed decision-making processes inflate the probability of unsound clinical hypotheses being accepted and subsequently used to generate interventions (Meehl, 1997).

The practice of subtest interpretation has long been suspect (McNemar, 1964). The present results are congruent with the current professional literature regarding subtest profiles as unreliable and invalid. That is, there is considerable evidence that subtest profiles do not demonstrate acceptable accuracy in discriminating among diagnostic groups and do not substantially covary with socially important academic and psychosocial outcomes (Hale & Green, 1995; Kavale & Forness, 1984; McDermott et al., 1992; Watkins, 2003). Given this lack of empirical support, subtest profiles should not be interpreted (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

References

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29,* 52–64.

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

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57,* 717–726.

Bracken, B. A., McCallum, R. S., & Crain, R. M. (1993). WISC–III subtest composite reliabilities and specificities: Interpretive aids [Special issue, Monograph, WISC–III series]. *Journal of Psychoeducational Assessment,* 22–34.

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children—Third Edition. *Psychological Assessment, 10,* 285–291.

Canivez, G. L., & Watkins, M. W. (1999). Long-term stability of the Wechsler Intelligence Scale for Children—Third Edition among demographic subgroups: Gender, race/identity, and age. *Journal of Psychoeducational Assessment, 17,* 300–313.

Canivez, G. L., & Watkins, M. W. (2001). Long-term stability of the Wechsler Intelligence Scale for Children—Third Edition among students with disabilities. *School Psychology Review, 30,* 361–376.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Davidow, J., & Levinson, E. M. (1993). Heuristic principles and cognitive bias in decision making: Implications for assessment in school psychology. *Psychology in the Schools, 30,* 351–361.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989, March). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice, 17,* 420–430.

Faust, D. (1990). Data integration in legal evaluations: Can clinicians

deliver on their premises? *Behavioral Sciences and the Law, 7,* 469–483.
Groth-Marnat, G. (1997). *Handbook of psychological assessment* (3rd ed.). New York: Wiley.
Hale, R. L., & Green, E. A. (1995). Intellectual evaluation. In L. A. Heiden & M. Hersen (Eds.), *Introduction to clinical psychology* (pp. 79–100). New York: Plenum.
Kamphaus, R. W., Petoskey, M. D., & Rowe, E. W. (2000). Current trends in psychological testing of children. *Professional Psychology: Research and Practice, 31,* 155–164.
Kaufman, A. S. (1994). *Intelligent testing with the WISC–III*. New York: Wiley.
Kaufman, A. S., & Lichtenberger, E. O. (2000). *Essentials of WISC–III and WPPSI–R assessment.* New York: Wiley.
Kavale, K. A., & Forness, S. R. (1984). A meta-analysis of the validity of Wechsler Scale profiles and recategorizations: Patterns or parodies? *Learning Disability Quarterly, 7,* 136–156.
Kavale, K. A., & Nye, C. (1985–1986). Parameters of learning disabilities in achievement, linguistic, neuropsychological, and social/behavioral domains. *Journal of Special Education, 19,* 443–458.
Livingston, R. B., Jennings, E., Reynolds, C. R., & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: High for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology, 18,* 487–507.
Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "Intelligent Testing" approach to the WISC–III. *School Psychology Quarterly, 12,* 197–234.
McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25,* 504–526.
McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19,* 871–882.
Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice, 4,* 91–98.
Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.
Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly, 15,* 376–385.
Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Author.
Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45,* 294–302.
Siegel, D. J., & Piotrowski, R. J. (1994). Reliability of WISC–III subtest composites. *Assessment, 1,* 249–253.
Thorndike, R. L. (1994). [Review of the book *Clinical assessment of children's intelligence*]. *Journal of Psychoeducational Assessment, 12,* 172–179.
Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion? *Scientific Review of Mental Health Practice, 2,* 118–141.
Wechsler, D. (1974). *Wechsler Intelligence Scale for Children—Revised.* New York: Psychological Corporation.
Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third Edition.* San Antonio, TX: Psychological Corporation.
Zeidner, M. (2001). Invited foreword and introduction. In J. J. W. Andrews, D. H. Saklofske, & H. L. Janzen (Eds.), *Handbook of psychoeducational assessment: Ability, achievement, and behavior in children* (pp. 1–9). New York: Academic Press.

Received March 19, 2003
Revision received June 16, 2003
Accepted August 1, 2003 ■