Psychology in the Schools, Vol. 34(3), 1997

CCC 0033-3085/97/030181-15

THE CALCULATION OF NEW AND IMPROVED WISC-III SUBTEST RELIABILITY, UNIQUENESS, AND GENERAL FACTOR CHARACTERISTIC INFORMATION THROUGH THE USE OF DATA SMOOTHING PROCEDURES

KEVIN S. MCGREW AND WADE WRIGHTSON

St. Cloud State University

Assessment practitioners are often encouraged to adopt an "intelligent" approach to the interpretation of intelligence tests. A fundamental assumption of the "intelligent testing" philosophy is that psychometric test information (e.g., subtest *g* loadings) should be considered during the interpretive process. The relevant psychometric information is provided in the form of sample-based estimates. Unfortunately, the accuracy of these estimates, and the subsequent qualitative classification of intelligence subtests (e.g., good, fair, poor), are influenced to an unknown degree by sampling error. The current study demonstrated how data smoothing procedures, procedures commonly used in the development of continuous test norms, can be used to provide better estimates of the reliability, uniqueness, and general factor characteristics for the WISC-III subtests. © 1997 John Wiley & Sons, Inc.

Assessment personnel are often encouraged to adopt an "intelligent" approach to the interpretation of intelligence test batteries. A fundamental assumption of the "intelligent testing" philosophy (Kaufman, 1979, 1994) is that psychometric test information (e.g., subtest *g* loadings) should be considered during the interpretive process. Such information includes (but is not limited to) knowledge of each subtest's reliability, validity, uniqueness, and general factor characteristics. This psychometric information is obtained through the application of appropriate statistical procedures (e.g., factor analysis) to age or grade-based subsamples of a test batteries standardization sample.

Although the subsamples upon which subtest psychometric information is calculated are typically large, these test statistics are *sample-based* estimates and thus, contain sampling error (McGrew, 1994). Diekhoff (1992, p. 87) states that "because no sample is likely to match the population from which it is drawn in every way, the samples can be said to be in error, thus the term sampling error." For example, the internal consistency reliabilities for the WISC-III Digit Span test (Wechsler, 1991) are plotted by age in Figure 1. Visual inspection of the sample-based reliability coefficients in Figure 1 reveals a systematic age-related increasing trend, but also variation or "bounce" between adjacent age groups. This variation most likely reflects sampling error. If it were possible to calculate the subtest reliabilities for the entire population at each age level, the plotted reliabilities would most likely be more consistent (i.e., display less age-to-age variation). The data presented in Figure 1 indicates that a sample-based reliability coefficient at a specific age may be a biased over- or underestimate of the true population reliability.

Test developers have long recognized the effect of sampling error on the means and standard deviations for test scores calculated from the age- or grade-based standardization subsamples. To deal with the resulting sampling bias, smoothing or curve fitting procedures (Goodall, 1990) are typically used to more accurately estimate the population means and standard deviations. The goal of smoothing is to separate the data into a smoothed component that approximates the population parameters and a rough component that represents sampling error (Goodall, 1990).

Within a series of related statistical estimates, smoothing techniques replace a particular sample estimate with a value that is based on some form of *averaging* of surrounding observations. These smoothed values, and not the biased sample estimates, are used to develop a test's final norm tables (Zachary & Gorsuch, 1985). These "continuous norming" based estimates are better estimates of the population parameters for each age group because they are derived from analyses of all subjects in

Correspondence should be sent to Kevin S. McGrew, St. Cloud State University, Department of Applied Psychology, 720 4th Ave. So., St. Cloud, MN 56301. E-mail: mcgrew@tigger.stcloud.msus.edu.

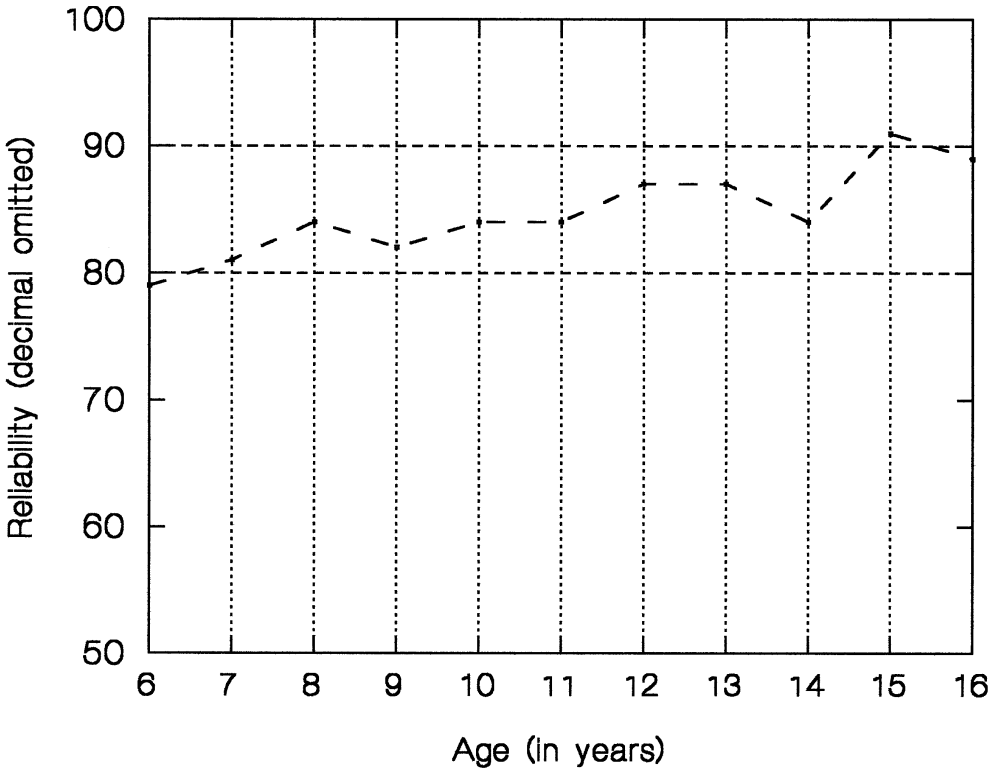

FIGURE 1. WISC-III Digit Span sample-based reliabilities by age.

the normative sample, not just from those of one age level (Zachary & Gorsuch, 1985). Although frequently applied to the development of a test's norms, test developers have not attempted to account for sampling error in the calculation and estimation of other important test characteristics.

Following the recommendation of Zachary and Gorsuch (1985), the purpose of this study was to demonstrate the application of data smoothing or curve-fitting procedures to the estimation of other psychometric information important in the interpretation of intelligence tests. More specifically, data smoothing techniques were applied to the calculation of three WISC-III subtest psychometric characteristics (i.e., general factor or *g* loadings, uniqueness, and reliability) that are often mentioned as critical to the test interpretation process (e.g., see Kaufman, 1979, 1990, 1994).

Briefly, subtest *g* loadings are typically based on each subtest's loading on the first unrotated factor or component in principal factor or component analyses. The *g* loading estimates are interpreted as reflecting each subtest's relationship to the theoretical construct of general intelligence (*g*) and are believed to be useful when evaluating whether a specific subtest deviation from a test battery profile may be clinically significant (Kaufman, 1994). Test *uniqueness* or specificity refers to the portion of a subtest's variance that is unique to the test (Kaufman, 1979, 1990). Uniqueness is used to determine the degree to which a subtest can be interpreted as measuring a unique ability that is not accounted for by abilities shared with other tests within an intelligence battery. Finally, *reliability* "refers to the degree to which test scores are free from errors of measurement" (APA, 1985, p. 19), and is thus important in determining to what extent a subtest's scores are accurate and constant and not due to random variation.

The purpose of this study was not to focus on the clinical relevance of subtest *g*, uniqueness, or

reliability when interpreting the WISC-III subtests. Rather, the purpose of this study was to demonstrate how better estimates of these WISC-III subtest technical characteristics can be obtained by applying data smoothing procedures to the original sample-based statistical estimates. New and improved estimates of the WISC-III subtest *g*, uniqueness, and reliability population parameters were calculated and compared to the original sample-based estimates of these technical characteristics.

Method

Sample

The data used in these analyses were the general factor or *g* loading, uniqueness, and reliability estimates for the 13 WISC-III tests based on the standardization sample. The WISC-III norm sample is a nationally representative sample that includes 2,200 individuals, with 200 at each of the 11 age groups, from ages 6 through 16 (Wechsler, 1991). The 11 age-based subsamples were used in all analyses. More detailed information regarding the complete sample can be found in the WISC-III manual (Wechsler, 1991).

Procedures

In a search of the literature, we did not find published general factor loadings for each of the WISC-III tests at each age level. Average values or ranges have been published (Kamphaus, 1993), but not values for each subtest at each age level. Therefore, we used the published WISC-III intercorrelation matrices (Wechsler, 1991) as input into a standard statistical computer program. At each age level, an individual subtest's loading on the first unrotated component in principal component analysis was used as the estimate of the test's general factor (*g*) loading (Kaufman, 1979; Sattler, 1992). The individual subtest reliabilities at each age were taken from Table 5.1 in the WISC-III manual. Previously calculated and published test uniqueness values for each subtest at each age level were taken from Bracken, McCallum, and Crain (1993). Since the classification of subtest uniqueness also requires the use of the error variance for each test, these values were calculated directly from the published WISC-III reliabilities (i.e., subtracting each test's reliability estimate from 1.0 to estimate the error variance for that subtest at a specific age).

The *g*, uniqueness, reliability, and error variance values were plotted by age for each subtest. The LOWESS (Cleveland, 1979, 1981) smoothing algorithm and the 2-D data plot option in the SYGRAPH (Wilkinson, 1990) graphic software package were used to smooth each test estimate across the 11 age groups. The SYGRAPH software was selected because it incorporates many automatic scaling, positioning, and plotting routines that produce accurate and clear graphs based on important design principles identified through cognitive science and graphic design research (Wainer, 1984; Wainer & Thissen, 1981; Wilkinson, 1990). The selection of the most appropriate data smoother for a set of data ultimately is left to the user because no generally satisfactory selection method has been developed (Goodall, 1990). A number of different SYGRAPH smoothing algorithms (e.g., LOWESS, splines, DWLS, quadratic smoothing) were applied to the WISC-III test characteristics. The LOWESS algorithm was selected for use in this study as visual inspection of the results from the different algorithms suggested that it provided the best solutions (i.e., smoothed curves that were not unduly influenced by values at specific ages). Also, of the available algorithms, LOWESS is "a flexible, effective smoother well-suited to everyday use for exploring data" (Goodall, 1990, p. 173). The LOWESS smoother produces a smoothed curve by running along the X values (i.e., WISC-III age groups) and finding a predicted value from a weighted average of nearby Y values (i.e., the respective test characteristic) (Wilkinson, 1990).

After each test characteristic was smoothed with the LOWESS algorithm, the respective original and smoothed values were both classified and then compared for differences at each respective

age level according to established qualitative categorical systems (Kaufman, 1979, 1990, 1994). In the case of *g* loadings, subtests were classified as *high* (.70 or above), *medium* (.51 to .69), or *low* (< .51). Uniqueness classifications were either ample, adequate, or inadequate (Cohen, 1959; Kaufman, 1979, 1990; Sattler, 1992). *High* uniqueness was defined as being present when an individual test's unique variance was equal to or above 25% of the total variance for the test and it exceeded the subtest's error variance. When a subtest's characteristics met only one of these criteria, it was considered to have *medium* uniqueness. When a subtest's characteristics meet neither of the criteria it was classified as *low.* Finally, based on the standards suggested by Salvia and Ysseldyke (1991), reliability estimates were classified as *high* (.90 or above), *medium* (.80 to .89), or *low* (< .80).

Results

An example of the smoothed reliability results for the WISC-III Digit Span test are presented in Figure 2.

A review of Figure 2 reveals that the smoothed curve reliability estimates for Digit Span follow the same systematic age-related increasing trend as do the original sample-based estimates, but with much less sampling error or "bounce". The most noticeable differences between the original and smoothed values were at ages 14 (.84 vs .88) and 15 (.91 vs .88). Although these differences are not extremely large, when using the previously defined low, moderate, and high reliability classification

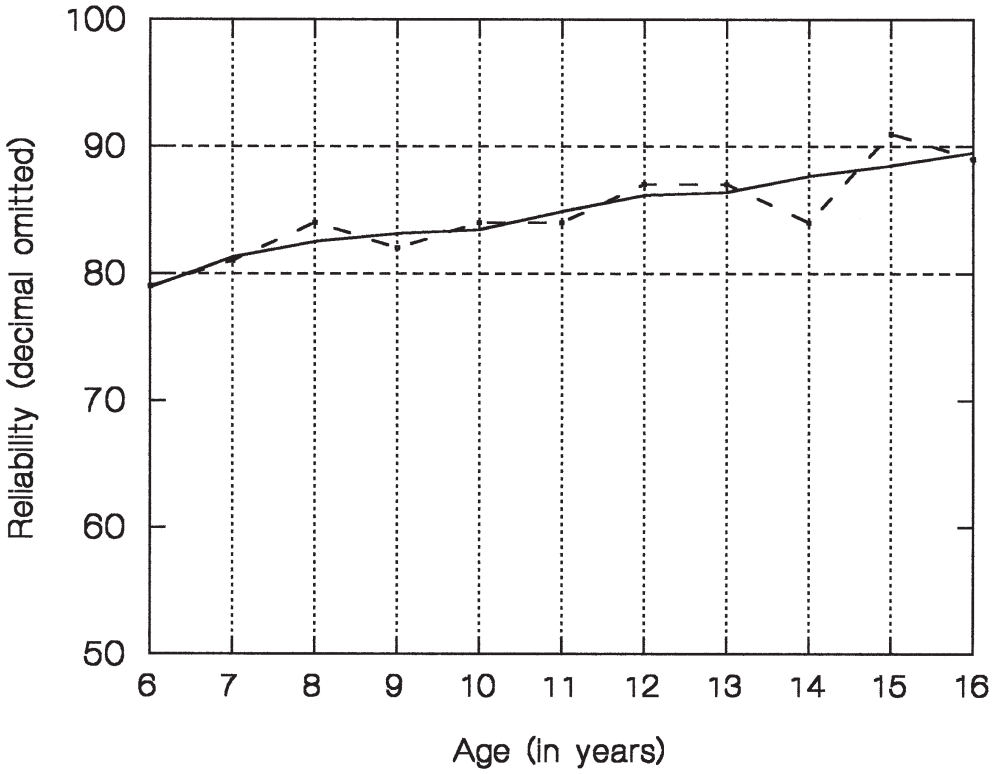

Figure 2. WISC-III Digit Span sample-based and smoothed reliabilities by age. The values connected by dotted lines represent the original sample-based reliability estimates (same as in Figure 1) for the Digit Span subtest. The solid line represents the LOWESS-based smoothed reliability estimates for Digit Span. The two horizontal dashed lines demarcate the three reliability categories (i.e., < .80 = low; ≥ .80 and < .90 = moderate; ≥ .90 = high).

system, the Digit Span's classification changed from the original sample-based classification of high to the smoothed classification of moderate at age 15. All original and smoothed reliability estimates for the 13 WISC-III subtests at each age group are summarized in Table 1.

A review of Table 1 indicates that of the possible 143 different reliability estimates (11 age groups × 13 subtests), 19 reliability classifications (13.3%) changed as a result of the smoothing process. Most of the changes occurred at ages 13 ($n = 4$) and 15 ($n = 5$). Another important observation is that smoothed reliability estimates are available for *all* 11 age groups for the two WISC-III speeded tests (i.e., Coding and Symbol Search). Because these speeded tests require special test-retest reliability subsamples and analyses (in contrast to the internal consistency estimates based on the available norm data), the test developers reported original values for only six age groups. The ability to estimate reliabilities for the missing age groups demonstrates another advantage of the use of data smoothing procedures in the estimation of reliability characteristics. The original and smoothed general factor loading and uniqueness estimates for all WISC-III subtests are presented in Tables 2 and 3. Sixteen (11.2%) of the *g* loading classifications changed as a result of the smoothing process (see Table 2). The largest number of classification changes ($n = 3$) occurred at ages 9, 11, and 15, and occurred most frequently ($n = 4$) for the Comprehension and Digit Span tests.

A review of Table 3 finds that the largest number ($n = 25$; 17.5%) of psychometric characteristic classifications changed for the subtest property of uniqueness. The largest number of changes occurred at ages 11 ($n = 5$) and 7 and 9 ($n = 4$), and were most frequent for the Similarities and Arithmetic subtests ($n = 5$), followed by Vocabulary and Block Design ($n = 4$). It is important to note that some of the changes in uniqueness classifications were due to changes in the smoothed estimates for either the subtest uniqueness or error variance (not reported in Table 3) or, in some cases, both estimates. This explains why a classification change may be indicated at a particular age for a test in Table 3 when there is very little or no difference in the reported uniqueness estimates (e.g., Arithmetic at age 9).

Table 4 summarizes the smoothed subtest reliability, uniqueness, and general factor loading classifications for all WISC-III subtests at all ages. This summary table is provided for easy use by clinicians. Across all three psychometric properties, most of the classification changes occurred at ages 15 ($n = 11$), 7 and 11 ($n = 9$), and 9 ($n = 8$). The subtests that demonstrated the largest number of changes were Arithmetic and Comprehension ($n = 8$) and Similarities ($n = 7$). The tests that demonstrated the fewest changes across psychometric classifications were Picture Arrangement and Mazes ($n = 1$), followed next by Picture Completion and Symbol Search ($n = 3$).

Discussion

Important decisions for individuals are often made on the basis of the interpretation of the subtest profiles of intelligence test batteries. To add as much objectivity as is possible to this clinical process, an "intelligent" approach to test interpretation is often recommended, an approach that involves the integration of important subtest psychometric test characteristics (e.g., general factor loading, uniqueness, reliability) into the interpretative process. This study demonstrated a potentially useful methodology (i.e., curve or data smoothing of sample-based psychometric test characteristics) for improving the estimation of important sample-based subtest psychometric properties.

Smoothed estimates of the 13 WISC-III subtest general factor loadings, uniqueness, and reliabilities across 11 age groups resulted in a 17.5%, 13.3%, and 11.2% change in the qualitative classifications. In many cases these qualitative classification changes were only due to minor differences in the absolute values of the respective subtest characteristic. For example, the original reliability of .81 and the smoothed value of .79 for the Arithmetic subtest at age 6 reflects only a .02 difference. However, the arbitrary cutting point between a medium and high reliability classification was a value of .80. This example, which also is relevant to the parallel classification systems used to catego-

Table 1
Comparison of Original and Smoothed WISC-III Subtest Reliabilities

Age group		Subtests												
		INF	SIM	AR	VOC	CMP	DS	PC	COD	PA	BD	OA	SS	MZ
6	O	73	81	81	82	75	79	78	75	82	82	71	69	80
	S	*73*	*80*	*79*	*82*	*73*	*79*	*79*	*73*	*83*	*81*	*70*	*71*	*80*
	D	—	01	02*	—	02	—	−01	02	−01	01	01	−02	—
7	O	76	77	73	79	72	81	84	70	84	77	65	76	78
	S	*76*	*80*	*77*	*85*	*77*	*81*	*81*	*73*	*80*	*81*	*67*	*73*	*78*
	D	—	−03*	−04	−06*	−05	—	03	−03	04	−04*	−02	03	—
8	O	86	84	78	88	85	84	81	—	72	83	65	—	76
	S	*79*	*81*	*75*	*88*	*78*	*82*	*82*	*75*	*75*	*82*	*68*	*74*	*74*
	D	07*	03	03	—	07*	02	−01	—	−03	01	−03	—	02
9	O	81	80	71	82	74	82	80	—	72	85	75	—	66
	S	*81*	*82*	*76*	*88*	*78*	*83*	*78*	*77*	*72*	*85*	*69*	*75*	*71*
	D	—	−02	−05	−06	−04	−01	02*	—	—	—	06	—	−05
10	O	82	82	79	88	79	84	74	78	74	89	69	72	70
	S	*82*	*82*	*77*	*88*	*76*	*83*	*76*	*78*	*72*	*85*	*69*	*76*	*68*
	D	—	—	02	—	03	01	−02	—	02	04	—	−04	02
11	O	85	82	79	88	76	84	76	82	70	84	65	79	68
	S	*84*	*82*	*78*	*88*	*78*	*85*	*74*	*79*	*74*	*86*	*67*	*76*	*68*
	D	01	—	01	—	−02	−01	02	03*	−04	−02	−02	03	—

12	O	85	84	74	89	81	87	72	—	79	87	68	—	66
	S	*85*	*82*	*78*	*89*	*77*	*86*	*73*	*81*	*75*	*87*	*69*	*77*	*68*
	D	—	02	−04	—	04*	01	−01	—	04	—	−01	—	−02
13	O	85	74	81	89	73	87	72	—	76	90	75	—	70
	S	*86*	*82*	*78*	*90*	*76*	*86*	*72*	*82*	*77*	*89*	*73*	*78*	*69*
	D	−01	−08*	03*	−01*	− 03	01	—	—	−01	01*	02	—	01
14	O	87	84	77	91	76	84	72	70	78	90	60	75	70
	S	*87*	*81*	*79*	*90*	*76*	*88*	*74*	*83*	*76*	*90*	*74*	*79*	*67*
	D	—	03	−02	01	—	−04	−02	−13*	02	—	−14	−04	03
15	O	88	81	81	91	80	91	82	90	73	92	76	82	61
	S	*88*	*83*	*80*	*90*	*76*	*88*	*75*	*86*	*75*	*91*	*74*	*80*	*66*
	D	—	−02	01	01	04*	03*	07*	04*	−02	01	02	02*	−05
16	O	88	84	82	89	73	89	75	—	73	90	71	—	67
	S	*88*	*84*	*82*	*89*	*74*	*89*	*75*	*89*	*73*	*91*	*71*	*81*	*66*
	D	—	—	—	—	−01	—	—	—	—	−01	—	—	01

Note. O = original, S = smoothed; D = difference between original and smoothed. INF = Information; SIM = Similarities; AR = Arithmetic; VOC = Vocabulary; CMP = Comprehension; DS = Digit span; PC = Picture completion; COD = Coding; PA = Picture arrangement; BD = Block design; OA = Object assembly; SS = Symbol search; MZ = Mazes.

*Changes in reliability classification (low, medium, high) due to smoothing. Decimals omitted from table.

Table 2
Comparison of Original and Smoothed WISC-III General Factor or g Loadings

Age group		Subtests												
		INF	SIM	AR	VOC	CMP	DS	PC	COD	PA	BD	OA	SS	MZ
6	O	71	72	75	72	62	59	69	31	68	76	69	65	55
	S	*71*	*72*	*74*	*71*	*60*	*59*	*68*	*30*	*69*	*75*	*68*	*65*	*51*
	D	—	—	01	01	02	—	01	01	−01	01	01	—	04
7	O	77	76	68	74	60	52	57	38	66	68	65	65	33
	S	*77*	*76*	*74*	*75*	*65*	*54*	*65*	*39*	*63*	*73*	*67*	*66*	*43*
	D	—	—	−06*	−01	−05	−02	−08	−01	03	−05*	−02	−01	−10
8	O	83	81	78	83	72	52	65	49	53	73	69	69	45
	S	*78*	*77*	*73*	*78*	*67*	*50*	*64*	*47*	*57*	*71*	*67*	*65*	*36*
	D	05	04	05	05	05*	02*	01	02	−04	02	02	04	09
9	O	79	79	70	78	65	44	69	51	54	70	66	62	25
	S	*79*	*79*	*73*	*80*	*70*	*50*	*67*	*50*	*56*	*70*	*66*	*63*	*35*
	D	—	—	−03	−02	−05*	−06*	02	01*	−02	—	—	−01	−10
10	O	76	78	71	82	75	54	66	50	61	69	63	59	41
	S	*76*	*79*	*72*	*79*	*70*	*50*	*67*	*49*	*55*	*69*	*64*	*59*	*32*
	D	—	−01	−01	03	05	04*	−01	01	06	—	−01	—	09
11	O	75	80	75	78	69	52	68	44	48	70	64	57	22
	S	*76*	*79*	*73*	*79*	*73*	*50*	*66*	*46*	*55*	*73*	*65*	*57*	*32*
	D	−01	01	02	−01	−04*	02*	02	−02	−07*	−03	−01	—	−10

12	O	81	80	70	79	78	43	64	44	60	78	71	57	33
	S	*76*	*79*	*73*	*79*	*73*	*47*	*65*	*44*	*55*	*76*	*64*	*56*	*32*
	D	05	01	−03	—	05	−04	−01	—	05	02	07*	01	01
13	O	76	76	73	80	70	48	63	45	55	80	61	52	41
	S	*78*	*78*	*73*	*80*	*72*	*46*	*65*	*43*	*57*	*78*	*62*	*57*	*33*
	D	−02	−02	—	—	−02	02	−02	02	−02	02	−01	−05	08
14	O	80	77	75	80	70	44	68	41	59	75	56	60	22
	S	*70*	*77*	*75*	*81*	*73*	*49*	*68*	*41*	*60*	*78*	*63*	*60*	*35*
	D	01	—	—	−01	−03	−05	—	—	−01	−03	−07	—	−13
15	O	80	76	74	85	80	58	74	54	67	80	74	66	48
	S	*79*	*77*	*76*	*80*	*74*	*51*	*66*	*40*	*63*	*78*	*65*	*64*	*37*
	D	01	−01	−02	05	06	07	08*	14*	04	02	09*	02	11
16	O	77	78	79	78	69	51	61	39	62	79	68	65	35
	S	*77*	*78*	*78*	*79*	*71*	*53*	*62*	*39*	*63*	*79*	*68*	*66*	*39*
	D	—	—	01	−01	−02*	−02	−01	—	−01	—	—	−01	−04

Note. O = original; S = smoothed; D = difference between original and smoothed. INF = Information; SIM = Similarities; AR = Arithmetic; VOC = Vocabulary; CMP = Comprehension; DS = Digit span; PC = Picture completion; COD = Coding; PA = Picture arrangement; BD = Block design; OA = Object assembly; SS = Symbol search; MZ = Mazes.

*Changes in *g* classification due to smoothing. Decimals omitted from table.

Table 3
Comparison of Original and Smoothed WISC-III Subtest Uniqueness Value

Age group		Subtests												
		INF	SIM	AR	VOC	CMP	DS	PC	COD	PA	BD	OA	SS	MZ
6	O	25	23	31	33	40	47	36	47	40	26	23	21	51
	S	*24*	*24*	*32*	*33*	*38*	*49*	*36*	*48*	*42*	*26*	*24*	*21*	*53*
	D	01*	−01	−01	—	02	−02	—	−01	−02	—	−01	—	−02
7	O	17	23	31	22	27	59	51	49	47	29	23	35	64
	S	*20*	*21*	*28*	*25*	*33*	*53*	*38*	*44*	*42*	*29*	*22*	*20*	*58*
	D	−03	02*	03	−03*	−06*	06	13	05	05	—	01	15*	06
8	O	17	16	22	20	31	52	40	36	37	31	17	18	58
	S	*18*	*19*	*27*	*20*	*31*	*58*	*40*	*39*	*43*	*32*	*23*	*18*	*59*
	D	−01*	−03	−05*	—	—	−06	—	−03	−06	−01	−06	—	−01
9	O	22	18	28	18	34	64	40	34	46	37	32	22	57
	S	*21*	*18*	*28*	*18*	*31*	*59*	*38*	*36*	*42*	*33*	*23*	*22*	*53*
	D	01	—	—*	—*	03	05	02	−02	04	04	09*	—*	04
10	O	23	18	32	13	26	60	35	40	40	30	17	29	44
	S	*24*	*18*	*28*	*19*	*29*	*59*	*35*	*41*	*43*	*33*	*24*	*29*	*53*
	D	−01	—	04	−06	−03	01	—	−01	−03	−03	−07	—	−09
11	O	28	19	21	26	28	52	32	51	44	33	26	38	61
	S	*23*	*19*	*26*	*20*	*25*	*60*	*32*	*47*	*44*	*30*	*19*	*35*	*52*
	D	05*	—*	−05*	06*	03	−08	—	04	—	03	07*	03	09

12	O	18	20	26	21	20	69	28	49	48	25	12	38	49
	S	*22*	*19*	*26*	*22*	*24*	*61*	*30*	*46*	*48*	*28*	*22*	*38*	*52*
	D	−04	01*	—	−01	−04	08	−02	03	—	−03	−10	—	−03
13	O	21	17	34	22	24	60	29	31	51	27	33	37	48
	S	*21*	*20*	*27*	*22*	*22*	*62*	*29*	*43*	*47*	*29*	*22*	*36*	*51*
	D	—	−03*	07	—	02	−02	—	−12	04	−02	11*	01	−03
14	O	23	24	20	23	23	60	30	41	42	35	17	32	59
	S	*21*	*21*	*26*	*21*	*22*	*60*	*30*	*40*	*41*	*31*	*24*	*34*	*48*
	D	02	03*	−06*	02*	01	—	—	01	01	04	−07	−02	11
15	O	17	20	26	16	19	60	28	44	30	29	25	35	38
	S	*21*	*22*	*24*	*21*	*22*	*61*	*33*	*45*	*37*	*31*	*23*	*31*	*44*
	D	−04	−02	02*	−05	−03	−01	−05	−01	−07	−02	02*	04	−06*
16	O	23	21	24	23	24	63	37	50	39	30	25	25	41
	S	*22*	*21*	*25*	*22*	*23*	*63*	*36*	*50*	*37*	*30*	*26*	*27*	*39*
	D	01	—	−01	01	01	—	01	—	02	—	−01	−02	02

Note. O = original; S = smoothed; D = difference between original and smoothed. INF = Information; SIM = Similarities; AR = Arithmetic; VOC = Vocabulary; CMP = Comprehension; DS = Digit span; PC = Picture completion; COD = Coding; PA = Picture arrangement; BD = Block design; OA = Object assembly; SS = Symbol search; MZ = Mazes.

*Changes in uniqueness classification due to smoothing of uniqueness and/or error variance (not reported in this table). Decimals omitted from table.

Table 4
WISC-III Test Reliability, Uniqueness, and General Factor Classifications Based on Smoothed Values

Age group		Subtests												
		INF	SIM	AR	VOC	CMP	DS	PC	COD	PA	BD	OA	SS	MZ
6	R	L	M	L*	M	L	L	L	L	M	M	L	L	M
	U	L*	M	H	H	H	H	H	H	H	H	L	L	H
	G	H	H	H	H	M	M	M	L	M	H	M	M	M
7	R	L	M*	L	M*	L	M	M	L	M	M*	L	L	L
	U	L	M*	H	H*	H*	H	H	H	H	H	L	L*	H
	G	H	H	H*	H	M	M	M	L	M	H*	M	M	L
8	R	L*	M	L	M	L*	M	M	L	L	M	L	L	L
	U	L*	L	H*	M	H	H	H	H	H	H	L	L	H
	G	H	H	H	H	M*	L*	M	L	M	H	M	M	L
9	R	M	M	L	M	L	M	L*	L	L	M	L	L	L
	U	M	L	H*	M*	H	H	H	H	H	H	L*	L*	H
	G	H	H	H	H	H*	L*	M	L*	M	H	M	M	L
10	R	M	M	L	M	L	M	L	L	L	M	L	L	L
	U	M	L	H	M	H	H	H	H	H	H	L	H	H
	G	H	H	H	H	H	L*	M	L*	M	M	M	M	L

11	R	M	M	L	M	L	M	L	L*	L	M	L	L	L
	U	M*	L*	H*	M*	H	H	H	H	H	H*	L	H	H
	G	H	H	H	H	H*	L*	M	L	M*	H	M	M	L
12	R	M	M	L	M	L*	M	L	M	L	M	L	L	L
	U	M	L*	H	M	M	H	H	H	H	H	L	H	H
	G	H	H	H	H	H	L	M	L	M	H	M*	M	L
13	R	M	M*	L*	H*	L	M	L	M	L	M*	L	L	L
	U	M	M*	H	M	L	H	H	H	H	H	L*	H	H
	G	H	H	H	H	H	L	M	L	M	H	M	M	L
14	R	M	M	L	H	L	M	L	M*	L	H	L	L	L
	U	M	M*	H*	H*	L	H	H	H	H	H	L	H	H
	G	H	H	H	H	H	L	M	L	M	H	M	M	L
15	R	M	M	M	H	L*	M*	L*	M*	L	H	L	M*	L
	U	M	M	M*	M	L	H	H	H	H	H	L*	H	H*
	G	H	H	H	H	H	M	M*	L*	M	H	M*	M	L
16	R	M	M	M	M	L	M	L	M	L	H	L	M	L
	U	M	M	M	M	L	H	H	H	H	H	M	H	H
	G	H	H	H	H	H*	M	M	L	M	H	M	M	L

Note. R = reliability; U = uniqueness; G = general factor loading; H = high; M = medium; L = low. INF = Information; SIM = Similarities; AR = Arithmetic; VOC = Vocabulary; CMP = Comprehension; DS = Digit span; PC = Picture completion; COD = Coding; PA = Picture arrangement; BD = Block design; OA = Object assembly; SS = Symbol search; MZ = Mazes.

*Change in classification due to smoothing.

rize subtest uniqueness and *g* characteristics, highlights a fundamental weakness of the traditional qualitative classification systems. If these qualitative classification systems are to be used, then values that are better estimates of the population parameters (i.e., smoothed estimates) should receive serious considerations by those who calculate and present this information for clinical use.

Because assessment professionals are often recommended to pay attention to the *g*, uniqueness, and reliability classifications of subtests when engaging in clinical interpretation, it is important that these classifications be as accurate as possible and be based on data that minimize the effect of sampling error. Smoothed estimates minimize the effect of sampling error. In addition, the presentation of the smoothed *g*, uniqueness, and reliability characteristics for each subtest in a graphic format similar to that represented by Figure 2 could be an alternative to the qualitative (e.g., high, medium, low) system. The visual-graphic presentation of smoothed subtest characteristics would allow clinicians the ability to quickly "see" the presence of any developmental trends in subtest characteristics and would more accurately portray how close a subtest with a particular classification was to the next highest or lowest classification category.

The smoothed WISC-III psychometric test estimates calculated for this study have three distinct advantages over the original sample-based estimates. First, the smoothed estimates are more likely better estimates of the population parameters. The data smoothing procedure reduced the effects of sampling error in the estimation of these important test characteristics. Second, the data smoothing procedures allow for the estimation of psychometric test characteristics at ages where data were not originally available. For example, the timed format of the WISC-III Coding and Symbol Search subtests make the calculation of internal consistency reliability estimates for these tests inappropriate. As a result of the additional expense and time involved in gathering test–retest data for timed tests, test developers typically conduct test–retest reliability studies at a smaller number of age group levels (e.g., six age groups for the WISC-III Coding and Symbol Search subtests). Through the use of data smoothing algorithms, this study was able to provide reasonable estimates of test–retest reliability estimates for all 11 age groups for the Coding and Symbol Search tests. By equally spacing test–retest studies for timed tests across the entire age range of a test it would be possible for test developers to provide reasonable interpolated reliability estimates for all ages. Third, similar to the process used to construct continuous norm tables (i.e., providing average scores for a test for each tenth of a school year when the norm data was only collected for every third of the school year), the use of data smoothing procedures makes it possible to provide estimates of important test characteristics at specific ages (e.g., 13 years 1 month) by "reading" values from the smoothed curves at any particular age or grade interval.

There are a number of limitations in this study that provide suggestions for future research. First, different smoothed paramater estimates might be obtained if a different smoothing algorithm was used (e.g., DWLS distance weighted least squares). Although the smoothed estimates might be different, it would be important to determine if the different estimates result in different qualitative (i.e., low, medium, high) classifications. Future research is needed to explore the impact on the results when different smoothing algorithms are used and, to determine whether certain "standard" algorithms might be identified and recommended for use in applied psychometrics. Second, the reading of smoothed values from curves is subjective and prone to error. A significant improvement would be to develop new software, or identify existing software packages, that can generate the exact smoothed paramater estimates for specific x/y coordinates along the final smoothed curves. Third, it must be remembered that the *g* loading and uniqueness estimates presented for the WISC-III subtests are only "within-battery" estimates. The subtest *g* loading and uniqueness estimates for a number of the WISC-III tests would most likely change if calculated on the basis of the WISC-III together with another intelligence test battery (McGrew & Flanagan, to appear; McGrew, Untiedt, & Flanagan, 1996).

The current study demonstrated the benefits of using data smoothing procedures in the estimation of important WISC-III subtest psychometric characteristics (i.e., reliability, uniqueness, *g* loadings). However, these procedures are not limited to intelligence test subtests or the three psychometric characteristics investigated. These procedures are equally applicable to tests in other domains (e.g., achievement, objective personality tests), to other test scores (e.g., composite scores such as the WISC-III Verbal, Performance, and Full Scale scores), and other test characteristics (e.g., standard errors, common factor loadings, etc.). The current study suggests that test developers should consider adding smoothed estimates of important psychometric test characteristics to the technical manuals for new or revised tests. Independent researchers who extract psychometric test characteristic information from test technical manuals in order to classify tests in books and articles are also encouraged to consider smoothing the values before making test classifications that might be used by clinicians.

References

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

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

Cleveland, W. S. (1979). Robust locally weight regression and smoothing scatterplots. *Journal of the American Statistical Association, 74,* 829–836.

Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician, 35,* 54.

Cohen, J. (1959). The factorial structure of the WISC at ages 7–6, 10–6, and 13–6. *Journal of Consulting Psychology, 23,* 285–299.

Diekhoff, G. (1992). *Statistics for the social and behavioral sciences: Univariate, bivariate, multivariate.* Dubuque, IA: Wm. C. Brown Publishers.

Goodall, C. (1990). A survey of smoothing techniques. In J. Fox & J. Scott Long (Eds.), *Modern methods of data analysis* (pp. 126–176). Newbury Park, CA: Sage.

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

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York: John Wiley and Sons.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn and Bacon.

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

McGrew, K. S., & Flanagan, D. P. (to appear). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment.* Boston, MA: Allyn and Bacon. Manuscript in preparation.

McGrew, K. S. (1994). *Clinical interpretation of the Woodcock–Johnson Tests of Cognitive Ability—Revised.* Boston, MA: Allyn and Bacon.

McGrew, K. S., Untiedt, S., & Flanagan, D. P. (1996). Uniqueness and general factor characteristics of the Kaufman Adolescent and Adult Intelligence Test. *Journal of Psychoeducational Assessment, 14,* 208–219.

Salvia, J., & Ysseldyke, J. (1991). *Assessment in special and remedial education* (5th ed). Boston: Houghton–Mifflin.

Sattler, J. (1992). *Assessment of children's intelligence and special abilities* (3rd ed.). San Diego, CA: Sattler.

Wainer, H. (1984). How to display data badly. *American Statistician, 38,* 137–147.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 32, pp. 191–241). Palo Alto, CA: Annual Reviews.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children—Third Edition.* San Antonio, TX: Psychological Corporation.

Wilkinson, L. (1990). *SYGRAPH: The system for graphics.* Evanston, IL: SYSTAT.

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology, 41,* 86–94.