DOCUMENT RESUME

ED 389 732 TM 024 306

AUTHOR Choi, Hee-sook
TITLE Scoring and Clerical Errors on the Stanford-Binet Intelligence Scale: Fourth Edition.
PUB DATE [91]
NOTE 18p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error Patterns; *Graduate Students; Higher Education; *Intelligence Tests; Scores; *Scoring; *Test Results
IDENTIFIERS Accuracy; Clerical Skills; *Stanford Binet Intelligence Scale Fourth Edition

ABSTRACT

Twenty-eight protocols of the Stanford-Binet Fourth Edition (SB:IV) obtained from graduate students were examined for scoring and clerical errors that contributed to the inaccuracy of test scores. Scoring of individual items was identified as the most error prone process, as evidenced by the fact that 96% of the protocols contained scoring errors. The most frequent scoring errors by subtest occurred on Comprehension, Vocabulary, Copying, Absurdities, and Verbal Relations. A relatively high occurrence of basal or ceiling errors (61%) was also found, with the Copying subtest being the most problematic for the establishment of both basal and ceiling levels. Clerical errors involving computation and coding were found in 32% of the protocols. Although the overall impact of examiner errors was fairly small in magnitude on the composite scores, 7% of the protocols produced a discrepancy of 6 points of more, sufficient to affect classification and placement decisions. (Contains 4 tables and 16 references.) (Author/SLD)

Scoring and Clerical Errors on the Stanford-Binet Intelligence Scale: Fourth Edition

Hee-sook Choi, Ph.D.
University of South Dakota

Running head: SCORING AND CLERICAL ERRORS ON SB:IV

Abstract

Twenty-eight protocols of the SB:IV obtained from graduate students were examined for scoring and clerical errors which contributed to the inaccuracy of test scores. Scoring of individual items was identified as the most error prone process as 96% of the protocols contained scoring errors. The most frequent scoring errors by subtest occurred on Comprehension, Vocabulary, Copying, Absurdities, and Verbal Relations. A relatively high occurrence of basal and/or ceiling errors (61%) was also found, with the Copying subtest being the most problematic for the establishment of both basal and ceiling levels. Clerical errors involving computation and coding were found in 32% of the protocols. Although the overall impact of examiner errors was fairly small in magnitude on the composite scores, 7% of the protocols produced a discrepancy of 6 points or more, sufficient enough to affect classification and placement decisions.

Scoring and Clerical Errors on the Stanford-Binet Intelligence Scale: Fourth Edition

The Stanford-Binet Intelligence Scale: Fourth Edition (SB:IV) is one of the best known and most widely used intellectual assessment measures, due in part to what are considered sound psychometric properties of the test. In the Technical Manual of the SB:IV (Thorndike, Hagen, & Sattler, 1986a), internal consistency coefficients for the composite are reported to range from .95 to .99 across age levels. Stability coefficients of preschool and elementary-school samples are .91 and .90, respectively, and the median standard error of measurement is 2.8 for the composite. The extensive reliability data provided in the manual appear to attest to the accuracy of IQ's obtained with the SB:IV. However, the reliability of the SB:IV is based on the spurious assumption that the test has been administered without any error on the part of the examiner. Neither the SB:IV nor any other major intelligence test includes interrater reliability coefficients associated with examiner errors. This is particularly problematic, considering that there are numerous studies with the Wechsler Intelligence Scales indicating frequent scoring and clerical errors made on the test protocols by both graduate students and professional psychologists (Slate & Hunnicutt, 1988; Bradley, Hanna, & Lucas, 1980; Levenson, Golden-Scaduto, Aiosa-Karpas, & Ward, 1988; Beasley, Lobasher, Henley & Smith, 1988; Slate & Chick, 1989; Sherrets, Gard, & Langner,1979: Miller & Chansky,1972: Blakey, Fantuzzo, Gorsuch, & Moon, 1987).

Beasley et al. (1988), for example, found that the difference between the scores trained psychologists calculated and the correct scores ranged from -11 to +13 for the Full Scale IQ on the Wechsler Intelligence Scale for Children- Revised (WISC-R). When taken into consideration that classification and placement decisions, especially in an educational setting, are frequently based on IQ test results as one of the key components, discrepancies of this magnitude could result in inappropriate placement or prevent a child from receiving needed special services. Moreover, the reliability of test scores is a prerequisite for the validity of a test. Examiner errors in scoring test protocols will decrease the reliability and, therefore, the validity of IQ's obtained.

Given the serious ramifications of examiner errors on intelligence tests, it is essential to investigate the nature and frequency of the errors as a first step to ensuring valid test scores. However, to date there has been a dearth of studies investigating examiner mistakes on the SB:IV which may contribute to the inaccuracy of test scores. This is at odds with the fact that the SB:IV is one of the most frequently recommended intelligence tests for use (Sattler, 1988), that the structure and administration of the SB:IV are substantially different from its predecessors (Glutting, 1989), and that scoring and computation of IQ's are well documented as an error-prone process (Conner & Woodall, 1983; Franklin, Stillman, Burpeau, & Sabers, 1982). The purpose of the present study was, therefore, to examine the frequency and types of scoring and clerical errors made by graduate students on the SB:IV protocols.

Method

Subjects

Over a 2 year time span, a SB:IV protocol was obtained from each of 28 graduate students enrolled in clinical and school psychology programs at the Master's and Specialist's levels, respectively. The students had completed an Individual Intelligence Testing course in which they were required to administer the SB:IV at least 6 times to child and/or adolescent volunteers. Each protocol was examined by the instructor or graduate assistant with verbal and written feedback given to the students. The final protocol and video-tape of administration for grading were submitted upon the seventh or eighth administration of the SB:IV for most of the students. The evaluation of the final protocols was conducted by the instructor. The 28 protocols examined for the purpose of this study were comprised of either the final protocols submitted for grading or the protocols that were obtained during a practicum course following completion of the intelligence testing class.

Procedure

A check list was designed for recording clerical and scoring errors noted during the examination of protocols. The errors examined included: 1) computational errors (test-age, summation of raw scores, and summation of Standard Age Scores), 2) coding errors (conversion of raw scores to Standard Age Scores and conversion of Standard Age Scores to Area Standard Age Scores), 3) errors in determining the entry level, 4) errors in establishing the basal and/or ceiling, 5) scoring errors, and 6) errors in questioning (failure to query and inappropriate query). Using the checklist, frequencies of occ urrence in error were obtained for the above categories. For each protocol, errors in a particular category were marked only once regardless of the number of times that particular error occurred. In addition, ambiguous responses that were clearly unscorable as pass or fail, using the Guide for Administering and Scoring for the SB:IV (Thorndike, Hagen, & Sattler, 1986b), were excluded from being in error.

Results

Table 1 presents the number of protocols which contained errors in each of the categories with their respective percentages. Nine protocols (32%) contained computation errors; that is, incorrect test-age (3%) and summation of raw scores (29%). The inaccurately computed birthdate was in error only by one month and, thus, did not alter the test scores. All protocols were error-free for the summation of Standard Age Scores (SAS's).

Insert Table 1 about here

Coding errors were found in nine protocols (32%), which were made when converting raw scores to SAS's (18%) and SAS's to Area SAS's (14%). These errors were due primarily to reading the wrong columns of scores in the SB:IV manual. In addition, five protocols (18%) contained errors in determining the entry level which were due either to incorrectly reading the Entry-Level Chart or to establishing an inappropriate ceiling on the Vocabulary subtest.

The majority of the protocols (61%) contained basal and/or ceiling errors. Table 2 further examines the three most frequently occurring subtests with errors in establishing either the basal or ceiling. Since the number of subtests administered varies depending upon the entry level of an examinee, the percentages weie computed based on the number of protocols with errors in conjunction with the total number of protocols which included that particular subtest. Both basal and ceiling errors were most frequently noted in the Copying subtest.

Insert Table 2 about here

Twenty-seven protocols (96%) contained scoring errors; that is, assigning credit to incorrect responses or failing to credit correct responses. Scoring errors were the most frequently occurring mistake in the protocols and the majority of them resulted from assigning credit to incorrect responses. Table 3 further delineates the subtests which contained errors in scoring. Errors were most frequent on those subtests requiring verbal responses among which Comprehension was the most error prone. In addition, the Copying subtest was identified as susceptible to substantial error as it ranked third in scoring errors.

Insert Table 3 about here

The next most frequent error was failure in questioning when necessary, which was noted in 26 protocols (93%). 14 protocols (50%) also contained inappropriate queries; that is, questioning when not appropriate. In general, students were more likely not to question when required than they were to unnecessarily question.

Table 4 shows the effects of clerical and scoring errors on the composite scores of the SB:IV. The differences between the students computed composite scores and the corrected composite scores ranged from -2 to +12. Although most of the differences were small and within one standard error of measurement (SEM) of the SB:IV composite, approximately 36% of the protocols contained discrepancies in excess of one SEM. Furthermore, in 7% of the protocols the composite scores computed by students were more than 5 points above the corrected values. It may be imperative to note that the largest difference of 12 points found in one protocol resulted, in large part, from coding errors of Area SAS's rather than scoring errors. Overall, the composite scores obtained by the students tended to be higher than the corrected values. Only 17.8% of the protocols contained composite scores which were lower than the corrected composite value with the magnitude of the discrepancies being 2 points or less.

Insert Table 4 about here

Discussion

The results of this study are consistent with the results of previous studies examining errors in computation and scoring on Wechsler Intelligence protocols (e.g., Beasley et. al, 1988; Slate & Chick, 1989). In the present study, simple clerical errors involving computation and coding were found in 32% of the protocols, confirming that examiners do frequently make mistakes in the process of transforming raw scores to IQ's. It also indicates that such errors are not confined to the Wechsler Scales and that psychologists using the SB:IV need to be cognizant of the relatively high incidence of such errors so as to prevent or minimize their occurrence.

Accurate scoring of individual items on the SB:IV was identified as the most error prone process. Almost all SB:IV protocols (96%) contained scoring errors, the majority of which resulted from crediting incorrect responses, thus inflating the test scores. It seems that graduate students were more apt to give credit when they were unsure of whether a response was acceptable.

The most frequent scoring errors by subtest occurred on Comprehension and Vocabulary, which is similar to previous findings concerning scoring difficulties with the Wechsler Intelligence Scales (Brannigan, 1975; Slate & Chick, 1989). A substantial number of the protocols also contained scoring errors on Copying, Absurdities, and Verbal Relations. These five subtests require subjective judgment on the part of examiners as no list of examples could be exhaustive of the possible responses an examinee might deliver. Therefore, in order to minimize scoring errors on these subtests, psychologists should be familiar with the expanded scoring criteria in the Guide for Administering and Scoring for the SB:IV. Too much reliance on the relatively few examples of common responses provided in the Item Books can significantly reduce scoring accuracy.

The accuracy of scoring was further impeded by either students' falling short of quering as specified in the manual or quering unnecessarily on responses that were clearly scorable as pass or fail. The students were nearly twice as likely not to query when necessary as they were to unnecessarily query. This tendency was also noted in the study with the WISC-R (Slate & Chick, 1989).

A considerably high occurrence of basal and/or ceiling errors (61%) was found in the protocols, suggesting that establishing a basal and/or ceiling is another oft-repeated error prone process. The problems were most frequently encountered on the Copying, Absurdities, Pattern Analysis, Vocabulary, and Quantitative subtests with Copying being the most problematic for both basal and ceiling levels. Difficulty with the establishment of basal and ceiling levels may be due, in part, to the inherent structure of the SB:IV in that for several subtests (e.g., Bead Memory, Quantitative, Pattern Analysis, and Copying) item types change with the introduction of new materials, directions, and sample items. This seems particularly troublesome in situations where items must be administered in reverse order to establish a basal level. Wersh and Thomas (1990) have also found that the determination of basal and ceiling levels on subtests which include item type changes is one of the problem areas in the administration of the SB:IV. Furthermore, there appear to be two conditions in which ceiling errors tend to be highly likely. First is that of premature discontinuance of subtest testing when 3 failures out 4 items in a row occur without ensuring that this condition occurred at two consecutive levels as specified in the manual. Second is where the examiner has the mistaken perception that all 4 items must be failed at two consecutive levels.

The overall impact of examiner errors on the composite of the SB:IV appears relatively minor, mostly resulting in small differences in the composite scores. However, 7% of the protocols produced a discrepancy of 6 points or more, sufficient enough to affect classification and placement decisions. It is, therefore, of utmost importance that psychologists make a conscious effort to ensure scoring accuracy as well as to eliminate careless clerical errors.

Further research concerning examiner errors on the SB:IV is needed to investigate whether the results from the present study can be generalized to applied settings. It is likely that practitioners working under time constraints and commonly with heavy caseloads may be more susceptible to making scoring and clerical errors than graduate students who are under constant supervision. Years of service may also be associated with lower ąccuracy on the part of practicing psychologists as many tend to rely on their memory for scoring.

References

Beasley, B. G., Lobasher, M., Henley, S., & Smith, I. (1988). Errors in computation of WISC and WISC-R intellience quotients from raw scores. Journal of Child Psychology and Psychiatry, 29(1), 101-104.

Blakeley, W., Fantuzzo, J., Gorsuch, R., & Moon, G. (1987). A peer-mediated, competency-based training package for administering and scoring the WAIS-R. Professional Psychology: Research and Practice, 18, 17-20.

Bradley, F., Hanna, G., & Lucas, B. (1980). The reliability of scoring the WISC-R. Journal of Consulting and Clinical Psychology, 48, 530-531.

Brannigan, G. (1975). Scoring difficulties on the Wechsler Intelligence Scale. Psychology in the Schools, 12, 313-314.

Conner, R., & Woodall, F. (1983). The effects of experience and structural feedback on WISC-R error rates made by student-examiners. Psychology in the Schools, 20, 376-379.

Franklin, M., Stillman, P., Burpeau, M., & Sabers, D. (1982). Examiner error in intelligence testing: Are you a source? Psychology in the Schools, 20, 376-379.

Glutting, J. (1989). Introduction to the structure and application of the Stanford-Binet Intellience Scale-Fourth Edition. Journal of School Psychology, 27, 69-80.

Levenson, R. L., Golden-Scaduto, C. J., Aiosa-Karpas, C. J., & Ward, A. W. (1988). Effects of examines' education and sex on presence and type of clerical errors made on WISC-R protocols. *Psychology Reports*, *62*, 659-664.

Miller, C., & Chansky, N. (1972). Psychologists' scoring of WISC protocols. *Psychology in the Schools*, *9*, 144-152.

Sattler, J. M. (1988). *Assessment of children*. San Diego: J. Sattler, Publisher.

Sherrets, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. *Psychology in the Schools*, *16*(4), 495-496.

Slate, J. R., & Chick, D. (1980). WISC-R examiner errors: Cause for concern. *Psychology in the Schools*, *26*(1), 78-84.

Slate, J. R., & Hunnicutt, L. C. (1988). Examiner errors on the Wechsler Scales. *Journal of Psychoeducational Assessment*, *6*, 280-288.

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

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986b). *Guide for administering and scoring*. Chicago: Riverside Publishing.

Wersh, J., & Thomas, M. R. (1990). The Stanford-Binet Intelligence Scale: Fourth Edition; observation, comments, and concerns. *Canadian Psychology*, *31*(2), 190-193.

Table 1. Type and Frequency of Errors in SB:IV Protocols

			Number of Protocols (N=28)	Percentage
I.	Computation			
	a)	Test-age	1	3%
	b)	Summation of Raw Scores	8	29%
	c)	Summation of SAS's	0	0%
II.	Coding			
	a)	Conversion of Raw Scores to SAS's	5	18%
	b)	Conversion of SAS's to Area SAS's	4	14%
III.	Entry Level		5	18%
IV.	Basal/Ceiling			
	a)	Basal	17	61%
	b)	Ceiling	17	61%
V.	Scoring of Items		27	96%
VI.	Questioning			
	a)	Failure to query	26	93%
	b)	Inappropriate query	14	50%

Table 2. Basal and Ceiling Errors in SB:IV Protocols

Subtests	Total Number of Protocols	N. with errors	Percentage
I. Basal			
Copying	12	7	58%
Absurdities	18	3	17%
Pattern Analysis	28	4	14%
II. Ceiling			
Copying	12	4	33%
Vocabulary	28	6	21%
Quantitative	28	6	21%

Table 3. SB:IV Subtests with Scoring Errors

Subtests	Total Number of Protocols	N. with errors	Percentage
Comprehension	28	27	96%
Vocabulary	28	16	57%
Copying	12	6	50%
Absurdities	18	5	28%
Verbal Relations	9	2	22%
Pattern Analysis	28	3	11%
Quantitative	28	3	11%
Memory for Objects	22	1	5%

Table 4. Magnitude and Direction of Discrepancies in IQ

Difference	Number of Protocols	Percentage
-2	1	3.5%
-1	4	14.3%
0	1	3.5%
+1	9	32.1%
+2	3	10.7%
+3	4	14.3%
+4	4	14.3%
+6	1	3.5%
+12	1	3.5%