University of Memphis
University of Memphis Digital Commons

Electronic Theses and Dissertations

12-1-2011

Exchangeability of Brief and Abbreviated Intelligence Tests: Illuminating the Influence on Error Variance Components on IQs

Sarah McCallum Irby

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation
Irby, Sarah McCallum, "Exchangeability of Brief and Abbreviated Intelligence Tests: Illuminating the Influence on Error Variance Components on IQs" (2011). *Electronic Theses and Dissertations*. 375.
https://digitalcommons.memphis.edu/etd/375

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.

EXCHANGEABILITY OF BRIEF AND ABBREVIATED INTELLIGENCE TESTS: ILLUMINATING THE INFLUENCE ON ERROR VARIANCE COMPONENTS ON IQS

by

Sarah McCallum Irby

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Psychology

The University of Memphis

December, 2011

Acknowledgements

I am grateful for my supervisor, Randy Floyd, whose encouragement, guidance, and support from the initial to the final level enabled me to develop an understanding of the subject. I would also like to thank Kevin Newton for all of his help during the data entry process. Lastly, I offer a special thanks to Abigail Brown, Hannah Gibert, Leigh Nevill, Rachel Peterman, and Adam Schepman for their commitment and dedication throughout the entire data collection process.

Abstract

Irby, Sarah McCallum. M.S. The University of Memphis. December, 2011. Exchangeability of brief and abbreviated intelligence tests: Illuminating the influence on error variance components on IQs. Major Professor: Randy G. Floyd.

This study examined the relations between and the exchangeability of IQs from four brief and abbreviated intelligence tests. All four tests were administered to 40 college students and scored by one set of examiners and later scored by a second examiner. All IQs were submitted to a Generalizability theory analysis to examine the relative contributions of error variance components of "test" and "examiner" and their interactions in producing variance in IQs relative to the object of measurement, individual differences in general intelligence. Despite very strong mean reliability coefficients (i.e., .91 to .96), the resulting dependability coefficient was .75, which indicated suspect dependability. The inadequate dependability coefficient from this study indicates that IQs are not as exchangeable as one might have assumed based on internal consistency reliability estimates, inter-rater reliability estimates, and convergent validity evidence.

Table of Contents

List of Tables

Exchangeability of Brief and Abbreviated Intelligence Tests: Illuminating the Influence on Error Variance Components on IQs

In a survey of clinical, counseling, and school psychology training program directors, Belter and Piotrowski (2001) found that intelligence tests are considered to be essential for practice.. In several other surveys (e.g., Camara, Nathan, & Puente, 2000; Rabin, Barr, & Burton, 2005; Ryba, Cooper, & Zapf, 2003; Wilson & Reschly, 1996), intelligence tests were identified as some of the most frequently used assessment tools in clinics, forensic settings, and schools. Their perceived necessity and frequency of use are related to the federal laws that require intelligence tests to be administered to determine eligibility for special education and if a diagnosis is warranted. An intelligence test is designed to measure an individual's cognitive abilities; general intelligence is a common source of individual differences found in an assortment of cognitive tasks (Jensen, 1993). However, each test is designed to do so in slightly different ways. In practice, these differences can be problematic because most psychologists administer one intelligence test and assume, with the exception of any glaring behavioral excesses or deficits displayed by the examinee, that the test yields a valid IQ for individuals. It is important to be mindful of several important issues related to IQs before accepting particular score as valid for making a diagnosis or placing a child in special education..

The relations between IQs from varying tests have been examined across hundreds of studies, but their exchangeability has only recently been targeted (Floyd, Clark, & Shadish, 2008). For the purposes of this study, exchangeability refers to the likelihood that IQs are the same despite the varying conditions under which they are obtained. In regard to test exchangeability, Floyd et al. found that about 25% of

individuals taking an intelligence test will obtain an IQ that is an average of 10 points higher or lower when compared to IQs from other tests. Little is known, however, about the reasons for this reduced exchangeability or the relative contributions of varying influences producing score differences. It is thought that characteristics of the test (including those stemming from norming, as investigated in Floyd et al., 2008) as well as differences due to the effects of examiners (see, for example, Ryan & Schnakenberg-Ott, 2003) are the two most probable influences on test scores.

Characteristics of the Tests

Random error. One characteristic that affects the exchangeability of intelligence tests is error in measurement (Bracken, 1987). Thus, no IQ is perfectly reliable and, due to random error, IQs will differ. For example, when two IQs are obtained, it is possible that one is producing a lower score within its range of hypothetical true scores and the other is producing a higher score within its range of hypothetical true scores, resulting in a large discrepancy between the two IQs and reducing the exchangeability of the IQs.

Test floors and ceilings. A second characteristic that affects exchangeability is the range of scores yielded by a test or its subtests. These ranges may be represented by the varying floor and ceiling levels, and they primarily affect scores for individuals who score at least two standard deviations above or below the mean (Bracken, 1987). For example, a bright adolescent may obtain a standard score of 128 on one intelligence test after yielding perfect scores on most every subtest yet obtain a score of 143 on another intelligence test with subtests with higher ceilings.

Normative samples. Other characteristics that affect test exchangeability are the recentness and representativeness of the normative sample. For example, the Flynn effect

is a product of the increase in the normative level of performance on intelligence tests scores over time (Flynn, 2006). Thus, when a new test is normed, those participating in the norm sample will perform better, on average, than a comparable sample of those who participated in earlier norm samples for previous editions of a test. The Flynn effect is examined by analyzing the mean differences between two tests to determine if there are significant differences between IQs from different tests. As a result, tests normed more recently will tend to produce lower norm-based scores for individuals than tests normed years before (McGrew, 2009a). Flynn (2006, 2009) focused on the normative sample of the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III; Wechsler, 1997) and asserted that likely these norms are not typical and possibly insufficient. Similarly, Floyd and colleagues (2008) found that the WAIS-III produced IQs that were notably higher than the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993), which was normed approximately 7 years earlier. In addition, the WAIS-III produced higher IQs on average than the Woodcock-Johnson III, Tests of Cognitive Abilities, General Intellectual Abilities (WJ III COG GIA; McGrew & Woodcock, 2001), which was normed approximately 4 years prior.

Modern day intelligence tests are normed to be representative of the U.S. population with respect to age, sex, race/ethnicity, geographic region, and socioeconomic level, but the question of the optimal size of the normative sample for each age level has been unanswered. Thus, it is unclear how many participants are needed to adequately represent the abilities of the target population. Most tests have smaller normative samples for the adult population. For example, the Stanford-Binet, Fifth Edition (SB-V; Roid, 2003) includes 1200 adults between the ages of 17 and 80+, in contrast to 3600 children

between the ages of 2 and 16. Additionally, some (Flynn, 2006, 2009) have alleged that the WAIS-III has norms that are "soft," so it tends to produce higher mean values than other tests normed at approximately the same time.

Content and response processes. The content and presentation of items within subtests within intelligence tests may vary substantially. These variations—and the examinee's interaction with them, with some performing better on some types of items and others performing well on other types of items—can result in varying scores across tests (McGrew 2009a, 2009b). For example, some tests require verbal responses to items, whereas other intelligence tests require few verbal responses and rapid motor responses. As a result, a child with strong verbal abilities may score higher on a subtest with lots of verbal items and fewer rapid motor response items but lower on a test with lots of rapid motor response items and fewer verbal items. Each of the characteristics previously mentioned (e.g., the Flynn effect, normative samples, and error in measurement), individually or in conjunction with other characteristics, can result in substantial differences in IQs across different tests and contribute to IQs that are not exchangeable.

Evaluation of Effects of Test Characteristics

Previous research examining the relations between intelligence tests includes correlational studies between two different tests, which are usually conducted as convergent or criterion-related validity studies. An example of a typical correlation is the moderate correlation between the Woodcock-Johnson III, Tests of Cognitive Abilities, General Intellectual Abilities (WJ III COG GIA; McGrew & Woodcock, 2001) and the WAIS-III (Wechsler, 1997; r = .67). Another example is the moderate correlation between the WJ III COG GIA and the KAIT (Kaufman & Kaufman, 1993; r = .75).

Other research examining the relations between intelligence tests have focused on the exchangeability of IQs from different tests. The exchangeability of IQs was examined by Floyd and colleagues (2008) across 7 intelligence tests and 6 samples in order to quantify the extent to which IQs differ on an absolute level. One analysis included IQs from three intelligence tests, and the other 10 analyses included IQs from two intelligence tests. Furthermore, Floyd et al. used Generalizability theory (Shavelson & Webb, 1991) to examine the magnitude of the effects of general characteristics of intelligence tests on variance in IQs.

For most comparisons, Floyd et al. (2008) found that the variance component associated with the test was negligible, contributing less than 4% of the variance for 5 of the 6 samples. However, in one sample targeting IQs from the WAIS-III and WJ III, the test contributed more than 25% of the total variance. The influence of the interaction between the individual and the test—and residual error—contributed sizable variance for all 6 samples (range = 7% to 27%). In fact, this variance component accounted for more than 20% of all variance in IQs in 9 of 11 IQ comparisons. Thus, the variance in IQs that is not due to individual differences in ability does not typically come from the test itself; instead, it in part comes from the individual's responses to test stimuli, task requirements, or response requirements or through subtle effects associated with variation in the representativeness of normative samples at different ages.

From this Generalizability theory analysis, a dependability coefficient was obtained; it provided an estimate of the exchangeability of the IQ based on the variance due to individual differences in general intelligence compared to all other variance components. For example, the Generalizability theory analysis employing IQs from the

WAIS-III, the KAIT, and the WJ III produced a dependability coefficient of .53. However, the remainder of the pairwise IQ comparisons across samples yielded somewhat higher dependability coefficients (M = .73) than the previously mentioned dependability coefficient for three tests. Additionally, these coefficient values were typically well below minimal standards for reliability as well as the internal consistency values for each IQ. Despite the evidence of minimal effects on IQs due to characteristics of the tests, per se, and some evidence of effects due to the variation across examinees in their response to those test characteristics, the Floyd et al. study is limited in that it examined only one class of error variance because the interaction between the individual and the test could not be separated from residual error. Furthermore, Floyd et al. did not evaluate the effects from examiners, thus leaving unaccounted variance that also could not be separated from the residual error.

Examiner Effects

Most of the studies, even those that are part of the normalization sample, examining IQ relations or exchangeability have not considered examiner effects explicitly (e.g., Floyd et al., 2008; Kaufman & Kaufman, 2004; McGrew & Woodcock, 2001; Roid, 2003; Wechsler, 1999). Numerous studies have, however, explored examiner effects in isolation, including inter-scorer agreement stemming from outright administration and scoring errors as well as from scoring subjectivity (e.g., Alfonso, Johnson, Patinella, & Rader, 1998; Erdodi, Richard, & Hopwood, 2009; Ryan & Schnakenberg-Ott, 2003; Slate, Jones, Coulter, & Covert, 1992). Their findings indicate that IQ scores may vary substantially depending on the number and types of errors made by examiners while administering and scoring the test as well as on the degree to which

examiners are affected by positive or negative decision frames when scoring items that require subjective judgments.

Administration and scoring errors. Examiners frequently commit errors while administering and scoring intelligence tests. According to APA, "psychologists retain responsibility for the appropriate application, interpretation, and use of assessment instruments whether they score and interpret such tests themselves, or use automated or other services" (Standard 9.09c, Ethics Code, 2002)., Regardless of the training level, however, psychologists appear to make a significant number of errors in intelligence testing. The most common errors include those involving summing of item scores, recording the wrong scaled scores (based on subtest raw scores) when referencing norms tables, and entering the wrong raw scores into the computer scoring software (Slate & Hunnicutt, 1988). Such errors can produce substantial score differences. For example, Ryan, Prifitera, and Powers (1983) have shown that such errors may change IQs on the tests like the Wechsler Adult Intelligence Scale, Revised Edition (WAIS-R; Wechsler, 1981) by 4 to 18 points.

As evident in Table 1, most of the research in this area has focused on errors demonstrated on the Wechsler tests, but the same errors appear to be evident when administering and scoring other intelligence tests as well. Three of these 10 studies listed in Table 1 revealed that protocols completed by psychologists and graduate students were found to have at least one error that led to a change in IQs on about 80% of the protocols (Erdodi et al., 2009; Hunnicutt, Slate, Gamble, & Wheeler, 1990; Slate et al., 1992). The most common errors across the 10 studies included failure to query, failure to record answers verbatim, calculating IQs incorrectly, failure to obtain true ceilings on subtests,

and incorrectly adding item scores on subtests (Alfonso et al., 1998; Moon, Blakey, Gorsuch, & Fantuzzo, 1991; Ramos, Alfonso, & Schermerhorn, 2009). Even though these errors may seem like relatively minor ones, this body of research shows clearly that they result in significant changes in IQs (Kuentzel, Hetterscheidt, & Barnett, 2011).

Scoring subjectivity. Another type of examiner effect is associated with subjectivity in scoring responses that may vary substantially in quality. In scoring subtests that required the examiner to make such subjective judgments about the quality of responses, examiners may score items differently (especially using more liberal or conservative decision frames) resulting in different IQs across examiners (Ryan et al., 1983; Slate & Chick, 1989). For example, several studies have shown that scores obtained by different professionals scoring the same test protocol may differ by 6 to 17 IQ points (Bradley, Hanna, & Lucas, 1980; Conner & Woodall, 1983; Miller & Chansky, 1972).

Evaluation of Effects of Examiners

Examiner effects are typically evaluated in terms of inter-scorer agreement and inter-rater reliability. Inter-scorer agreement focuses on the item-score-by-item-score correspondence across at least one pair of examiners. It is typically reported as a percentage that stems from considering the proportion of matching item scores (i.e., agreements) to all possible items. Inter-rater reliability focuses on the relation between sums of items scores, such as raw scores or norm-based scores, which are continuous variables. It is typically reported as a Pearson product-moment correlation coefficient across scores from a pair of raters. Inter-rater reliability provides a more holistic

understanding of examiner effects on the relationship of different IQs versus inter-scorer agreement, which provides only a partial understanding of examiner effects.

As evident in Table 2, all three of the published studies examining inter-scorer agreement of IQs have focused on scoring of only the Verbal subtests from the Wechsler scales, which use a 3-point scale (e.g., 0, 1, and 2 points) based on sample responses and general criteria (e.g., degree of abstraction) shown in the manuals. These studies showed that, as a result of differences in how these Verbal subtests were scored, the IQs could vary by 4 to 18 points based on who was scoring the protocols (Bradley et al., 1980; Ryan et al., 1983; Ryan & Schnakenberg-Ott, 2003). Due to examiner errors on the Verbal subtests, these three studies indicate that there is only a 26% to 35% overall agreement in IQs. Therefore, because the current research (e.g., Bradley et al., 1980; Ryan et al., 1983) has focused on Wechsler tests, it is difficult to know to what extent scores of other subtests are affected by scoring subjectivity and to what extent the overall IQ is affected by the subjectivity of examiners.

Despite the above mentioned differences in IQs due to administration and scoring errors and scoring subjectivity, inter-scorer agreement, their total effects on IQ exchangeability, and their interactions with intelligence tests as a whole have yet to be evaluated thoroughly. However, some test manuals report inter-rater reliability for select subtests that may be affected by examiner subjectivity. It is likely that inter-rater reliability has not been evaluated for all subtests because there has yet to be an appropriate way to examine agreement or reliability of examiners aside from providing examiners with a protocol of responses to score (e.g., verbatim responses). Furthermore,

studies focusing on inter-scorer agreement have been limited in their scope of mainstream intelligence tests and need to expand their focus to include other current prominent tests.

Brief and Abbreviated Tests

Despite clear evidence that tests and examiners may have substantial effects on IQ exchangeability, recent studies have looked at only one of these influences at a time and have yet to examine both convergently. Ideally, both types of influences would be investigated through (a) one examiner administering multiple, full-length intelligence tests in one sitting (highlighting the test characteristics) and obtaining all IQs and (b) another examiner evaluating and scoring these administrations independently and obtaining IQs. However, most full-length intelligence tests can take anywhere from 45 minutes to several hours to administer. Thus, it is difficult to give more than two intelligence tests without separating them into multiple sessions, which can lead to possible confounds including participant attrition and events occurring between testing sessions affecting an examinee's responses (a.k.a., the confound of history). These problems can be overcome by using brief and abbreviated intelligence tests as proxies for their full-length counterparts. Brief and abbreviated tests can be administered in a short period, yet they yield IQs supported by substantial reliability and validity evidence (Homack & Reynolds, 2007).

An intelligence test is considered brief or abbreviated based on the amount of time required to administer the test as well as the information that is covered (Homack & Reynolds, 2007). Thus, brief and abbreviated intelligence tests can be completed quickly and their subtests sample from multiple specific cognitive ability domains to produce its scores. However, there is a slight difference between abbreviated and brief intelligence

tests. A brief intelligence test is a stand-alone test composed of only a few subtests, which are scored in reference to its own independent norm sample. Alternately, an abbreviated intelligence test is composed of select subtests from a full-length intelligence test that yields scores based on the norm sample from the full-length intelligence test.

The first abbreviated intelligence test was developed by Terman and Merrill (1937). This test could save the examiner about one-third of the testing time by omitting items from the Stanford-Binet Forms L and M (Terman & Merrill, 1937), yet it yielded a reliable IQ. The WAIS-R also has many abbreviated forms, which were based on different referral concerns and were composed of anywhere from 2 to 7 subtests out of the 10 core subtests (Ward & Ryan, 1996). Eventually, brief tests were developed for almost every test battery and were based on the idea of administering an abbreviated form of a longer battery (Kaufman, Kaufman, Balgopal, & McLean, 1996). However, abbreviated forms were much more common. One of the first brief tests to be well normed with a national sample was the Kaufman Brief Intelligence Test (K-BIT; Kaufman, & Kaufman, 1990), which included only two subtests. The K-BIT had a relatively large normative sample (N = 2,022) and took only about 30 minutes to administer.

Both brief and abbreviated intelligence tests continue to be widely used in practice today. Brief intelligence tests include the Wechsler Abbreviated Scales of Intelligence (WASI; Wechsler, 1999) and the Kaufman Brief Intelligence Test, Second Edition (KBIT-2; Kaufman & Kaufman, 2004), and abbreviated intelligence tests include those derived from the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG; McGrew & Woodcock, 2001) and the Stanford Binet, Fifth Edition (SB-V; Roid, 2003).

See Table 3 for a description of studies of the relations between IQs obtained from brief and abbreviated intelligence tests and IQs from other intelligence tests.

Purpose of the Study

Because they do not require extreme costs in time and effort, brief and abbreviated intelligence tests make it possible to examine the extent to which test properties and the effects from examiners influence the exchangeability of IQs.. It is common for test authors and other researchers to evaluate the effects of test properties across tests by examining the correlations between IQs. However, investigations of the effects from examiners on IQs have focused on Wechsler tests and have minimally evaluated examiner effects on other tests. To determine their total effect on IQ exchangeability as well the strength of these effects, it is important that both of these influences be examined in one study.

As an extension of the Floyd et al. (2008) study, this study examined the relations between and the exchangeability of IQs from four brief intelligence tests. All four tests were administered and scored by one set of primary examiners and also scored independently by a single secondary examiner. First, traditional analyses using correlations and tests of mean differences were completed to examine (a) the relations between and mean differences in IQs across tests and (b) the relations between mean differences in IQs produced by different scorers Then, all IQs were submitted to a Generalizability theory analysis to examine the relative contributions of error variance components of “test” and “examiner” and their interactions in producing variance in IQs relative to the object of measurement, individual differences in general intelligence.

Based on results from Floyd et al. (2008) and the Flynn effect (Flynn, 2006), it is hypothesized that more recent intelligence tests will yield lower IQs than those with earlier publication dates, which indicates that the test component will contribute to small but notable variance in IQs. Based on research focusing on the effects of scoring subjectivity, it is hypothesized that the WASI and SB-V would display lower correlations across examiners than the other intelligence tests because they require more subjectivity in scoring. As a result, the examiner-by-test interaction component would be expected to be sizeable. Based on findings from Floyd et al., it is hypothesized that the test-by-examinee interaction component would contribute the largest variance in IQs of any error variance component.

Method

Participants

Participants were 40 students attending a local university in the mid-south region of the United States. Students were selected from the Psychology Department's subject pool. Due to errors in using recording equipment, complete data for one participant were not collected, thus the participant was excluded from the analysis (n = 39). Approximately 80% were women, and 45% were White, 43% were Black, and 12% were otherwise classified (as Asian/Pacific Islander, Hispanic, Arab American, or Multiracial). Participants ranged in age from 18 to 46 with a mean age of 24.3 years (SD = 7.73), and 46.2% were freshmen, 23.1% were sophomores, 10.3% were juniors, and 20.5% were seniors. Participants were not excluded based on prior diagnosis or for any medications they were currently taking. Two participants reported a prior diagnosis of a mental health

disorder, and one participant reported currently being prescribed psychotropic medication.

Measures

Kaufman Brief Intelligence Test, Second Edition (KBIT-2). The KBIT-2 (Kaufman & Kaufman, 2004) is an individually administered brief test of intelligence designed for individuals ages 4 to 90 years. It consists of three subtests. The Verbal Knowledge subtest measures receptive vocabulary and general knowledge. The Riddles subtest measures verbal comprehension, reasoning, and vocabulary knowledge. The Matrices subtest measures the ability to solve new problems, perceive relationships, and complete visual analogies without testing vocabulary or language skill. Subtests yield raw scores that are converted into scaled scores for the three subtests and standard scores for the Verbal, Nonverbal, and IQ Composite scores. The IQ Composite has a mean of 100 and a standard deviation of 15.

According to Kaufman and Kaufman (2004), the KBIT-2 IQ Composite score has high internal-consistency reliability across adult samples ages 19 to 90 (mean split-half reliability coefficient = .95). The IQ Composite also has a high test–retest reliability (with an interval of 6 to 56 days between administrations) for the 13 to 21 and 22 to 59 age ranges (r = .92 and .90, respectively; Kaufman & Kaufman, 2004). Inter-rater reliability was not reported by the authors. The IQ Composite has satisfactory criterion-related validity based on correlations with other measures of brief and full-length intelligence tests as shown in Table 3.

Wechsler Abbreviated Scale of Intelligence (WASI). The WASI (Wechsler, 1999) is an individually administered brief intelligence test designed for individuals ages

6 to 89 years, and it consists of four subtests. The Vocabulary subtest requires the participant to define increasingly difficult vocabulary words. The Similarities subtest requires the participant to determine how two things are alike. The Block Design subtest requires the participant to assemble designs using up to nine blocks. The Matrix Reasoning subtest requires the participant to solve puzzles. From these subtests, the WASI yields two Full Scale IQs, one based on all four subtests, the FSIQ-4, and the other based on only two subtests (i.e., Vocabulary and Matrix Reasoning), the FSIQ-2. The mean of both FSIQs is 100 with a standard deviation of 15.

According to Wechsler (1999), both the FSIQ-4 and the FSIQ-2 have high internal consistency in the overall adult sample (mean reliability coefficient = .98 and .96, respectively). The test–retest reliability (with an interval of 2 to 12 weeks between administrations) for the overall adult sample for the FSIQ-4 and FSIQ-2 are high as well (r = .93 and .89, respectively). The inter-rater reliability was reported only for the Vocabulary and Similarities subtests, but both coefficients were very high (r = .98 and .99, respectively; Wechsler, 1999). As shown in Table 3, both the FSIQ-4 and FSIQ-2 have satisfactory criterion-related validity based on acceptable correlations with various intelligence tests. For the purposes of this study, only the FSIQ-2 was obtained because its reliability and validity evidence are reasonably strong and because one of the subtests contributing to only the FSIQ 4 (Block Design) would pose particular problems in scoring via the video and audio recording used in this study.

Woodcock-Johnson III Brief Intellectual Ability (BIA). The Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG; McGrew & Woodcock, 2001) is an individually administered full-length intelligence test designed for individuals ages 2

to 90+ years. The WJ III COG includes Brief Intellectual Ability (BIA) battery formed from three subtests. The Verbal Comprehension subtest measures several aspects of language development. The Concept Formation subtest measures the ability to discover the concept or rule that underlies a problem or set of images. The Visual Matching subtest measures processing speed. Subtests yield raw scores that are entered into computer scoring software and converted into the BIA standard score. The BIA score has a mean of 100 with a standard deviation of 15.

According the McGrew and Woodcock (2001), the BIA has demonstrated high internal consistency reliability for ages 18 to 59 (median reliability coefficient across ages 18 to 59 = .96). Test–retest and inter-rater reliabilities were not reported for the BIA, but test–retest reliability (with an interval of one day between administrations) was reported for the Visual Matching test for ages 26 to79 (r = .70). The BIA has fair criterion-related validity based on correlations with full-length test batteries as shown in Table 3.

Stanford-Binet Intelligence Scales (SB-V) Abbreviated Battery IQ (ABIQ). The Stanford-Binet Intelligence Scales, Fifth Edition (SB-V; Roid, 2003) is an individually administered full-length intelligence test designed for individuals ages 2 to 85+. The SB-V yields an Abbreviated Battery IQ (ABIQ) from performance on two routing subtests. The Matrices subtest measures the participant's ability to solve new problems, perceive relationships and complete visual analogies without testing vocabulary or language skill. The Vocabulary subtest measures word knowledge and language abilities. Subtests yield raw scores that are converted into the ABIQ standard score. The ABIQ has a mean of 100 and a standard deviation of 15.

According to Roid (2003), the ABIQ has high internal consistency reliability for the overall sample (mean reliability coefficient = .91). The overall test–retest reliability (with an interval of 5 to 8 days between administrations) for the ABIQ was moderately high for both the 6 to 20 and 21 to 59 age groups (r = .84 and .80, respectively). The inter-rater reliability was not reported for the ABIQ, but it was reported for the Knowledge subtest for three pairs of examiners (range of r = .95 to .98). The criterion validity of the ABIQ was found to be .87 for ages 6 and above when correlated with the SB-V FSIQ (Roid, 2003).

Procedures

Policies and procedures dictated by the Institutional Review Board were adhered to during the participant recruitment and data collection.

Recruitment. Participants were recruited through the Psychology Department's subject-pool during their Introduction to Psychology course. The participants received credit in their Introduction to Psychology course for their participation in the study.

Testing sessions. Students registered for testing sessions online and were asked to arrive in the university's sponsored psychology clinic and check-in with the receptionist. Each testing session was completed in a quiet room in the clinic. Each test session was recorded and scored by the author in order to evaluate the inter-rater reliability of each brief intelligence test's scores. The rooms in the clinic where the testing took place had audio and video equipment already installed. In order to ensure that subtests that involve visual stimuli or manipulatives would be scored accurately by the author, a second video camera was aimed at the testing table to record these responses.

Participants provided written consent to participate in the study and completed a demographics form (see Appendix A). They completed the four brief intelligence tests in an approximately 2-hour session with no planned breaks. The tests were administered in a counterbalanced order using a Latin Square design. A one-way ANOVA was used to evaluate possible order effects on the IQs from the intelligence tests administered first, second, third, and fourth—regardless of which test it was. Results were nonsignificant ($p > .05$); thus, counterbalancing eliminated any order effects that may have been present.

Primary examiners. The brief intelligence tests were administered by five examiners (i.e., "primary examiners"). The primary examiners were advanced graduate students; four were education specialist students, and one was a doctoral student. The primary examiners were required to have passed two graduate-level assessment courses and a graduate-level practicum as well as to have completed two one-hour training sessions prior to administering any tests.

Each primary examiner completed a demographics survey (see Appendix B). The primary examiners reported completing 25 to 92 hours of graduate coursework with a mean of 42.4 hours prior to this study. These examiners also reported completing 200 to 1600 graduate practicum hours, with a mean of 523.2 hours prior to administering the tests. They reported administering 13 to 546 comprehensive and screening tests, with a mean of 109 tests. All primary examiners were White. Four primary examiners were female, and one was a male.

Training sessions consisted of reviewing administration and scoring procedures for the four different tests they would be administering as well as instruction in how to use the different electronic equipment. After the first training session, each primary

examiner submitted protocols for each test in order to ensure his or her competence in administration and scoring. The protocols were reviewed by the author to ensure that no invalidating errors were present, and the minor errors that were committed were discussed with the primary examiners during the second training session. Each primary examiner demonstrated competency on each battery prior to data collection.

Each primary examiner completed 8 assessments to promote an equal balance of contributions to the data set across examiners. After completing the testing sessions, the primary examiners were asked to complete the scoring of the instruments within 1 week. The primary examiners hand scored the protocols and used norms tables to arrive at norm-based scores the KBIT-2, WASI, and SB-V and used Woodcock Johnson III Normative Update Compuscore and Profiles (Woodcock, McGrew, & Mather, 2006) scoring software to produce norm-based scores for the WJ III COG. Primary examiners were asked to consult with other student examiners and school psychologists in the field if they had questions about scoring items from the brief intelligence tests. However, they were asked not to consult with the author (or faculty supervisor), who remained blind to the results from the tests and reviewed only the recordings of these sessions months later. After each administration, primary examiners placed completed protocols in folders in a filing cabinet monitored by the faculty advisor to ensure the secondary examiner remained blind before scoring.

After completion of each test session, each primary examiner completed a post-testing integrity checklist. The checklist included 5 questions. For the first question, 100% of examiners reported that they had administered the tests as they would in practice. For the second question, 97% reported administering all tests in one sitting.

Only one examiner did not administer all tests in one sitting because one subtest from the KBIT-2 was administered one week later after initially being omitted in error. For the third question, 95% reported administering the tests in the prescribed order. A total of two examiners did not administer the tests in the prescribed order because an incorrect Stanford-Binet protocol was included in the participant folder, and the Stanford-Binet was administered at the end of the test session. For the fourth question, 100% reported not consulting with the author or faculty supervisor regarding scoring or administration after the training was completed. For the fifth question, 95% reported they did not consult with others regarding scoring or administration. Only one examiner reported consulting with another examiner for two different participants.

Secondary Examiner. The author served as the secondary examiner and reviewed the video recordings of the sessions in order to score each test using new protocols. The secondary examiner was able to rewind and review responses multiple times for long verbal responses on different subtests (e.g., Vocabulary on the WASI) and when unsure of how to score responses. Several scoring issues were discussed with the faculty advisor, including what to do for inaudible responses or administration errors. For cases in which responses were inaudible or it was apparent that the primary examiner demonstrated an administration error (e.g., spoiling a response, not prompting or querying appropriately, and failing to establish a floor or ceiling), a list of random numbers was consulted to determine whether or not to award credit for items that were affected. For example, if the primary examiner skipped an item he or she was supposed to administer, the secondary examiner would review the random numbers list that was generated in order to decide whether to award credit. If an even number was selected, the

examiner awarded credit, and if an odd number was selected, the examiner did not award credit. The types of common errors made by the primary examiner that led the secondary examiner to consult the random numbers list are listed in Table 4.

Results

Data Screening and Tests of Assumptions

Preliminary data analyses were conducted with each of the four tests for each examiner to ensure that the assumptions of multivariate analysis and correlations were not violated (Tabachnick & Fidell, 2006). There were no missing values and no univariate (zs < $\left|2.7\right|$) or multivariate outliers found for any of these variables. No IQ was notably skewed for either examiner (all values < $\left|1.0\right|$). No IQ had notable kurtosis for either examiner (all values < $\left|1.0\right|$ except for the KBIT-2 for the secondary examiner; kurtosis = -1.02). The assumption of linearity was assessed using bivariate scatterplots and met. All assumptions of paired-samples *t*-tests were met.

Table 5 includes the means and standard deviations for each IQ by examiner. The means ranged from 97.03 (KBIT-2) to 103.10 (WASI) for the primary examiner and from 97.13 (KBIT-2) to 103.59 (WASI) for the secondary examiner. The means for both examiners were very close to 100. The standard deviations ranged from 9.03 (SB-V) to 11.97 (KBIT-2) for the primary examiner and from 8.37 (SB-V) to 11.37 (KBIT-2) for the secondary examiner. The standard deviations for both examiners were less than 15 in every case, which indicates restriction of range of the samples.

Convergent Validity and Mean Differences across Tests for Brief and Abbreviated IQs

To examine the convergent validity evidence supporting the IQs, correlations for each test with the other three tests for both examiners were conducted, resulting in 12 correlations. In instances of restriction or expansion of range in the IQs, the correlation coefficients were corrected for such error using the Incidental Variable correction from Attenuation correction 2.1 (Barrett, 2002). Table 6 includes both the uncorrected and corrected correlations between each of the tests' IQs as produced by each set of examiners. The following general labels were used for this study: *negligible,* .00 to .19; *weak,* .20 to 39; *moderate,* .40 to .69; *strong,* .70 to .89; and *very strong,* .90 to 1.0. Based on the labels mentioned, there is only one strong correlation for the primary examiners (between KBIT-2 and WASI). There are two moderate correlations (between the WJ III COG and the WASI and KBIT-2), two weak correlations (between the SB-V and the KBIT-2 and WJ III COG), and one negligible correlation (between the SB-V and the WASI). Similar results were seen for the secondary examiner. Only one correlation was strong (KBIT-2 and WASI), and the remaining included one moderate correlation (between the WJ III COG and the KBIT-2), two weak correlations (between the WASI and the WJ III COG and the SB-V), and two negligible correlations (between the SB-V and the WJ III COG and the KBIT-2).

Table 6 also includes correlations corrected for range restriction that represent the relations between each of the IQs as produced by each set of examiners. Based on the labels mentioned in Table 6, there are three strong correlations for the primary examiner. There were three moderate correlations (between the SB-V and each of the other tests)

and no weak or negligible correlations. Similarly, there were two strong correlations and four moderate correlations (including 3 between the SB-V and each of the other tests) for the secondary examiner. The most salient differences when corrected were noted for the SB-V, which had no moderate uncorrected correlations for either examiner.

A one-way ANOVA was conducted to compare the effects of each test on mean IQ for each examiner. For the primary examiners, there was a significant effect of test on IQs, $F(3, 152) = 3.43$, $p = .02$. For the secondary examiner, there was also a significant effect of test on IQs, $F(3, 152) = 3.29$, $p = .02$. Tukey post-hoc results indicated that there were significant differences between the WASI and the KBIT-2 scores and the WASI and the SB-V scores ($p < .05$) for both examiners. All other comparisons were nonsignificant ($p > .25$).

Inter-rater Reliability and Mean Differences across Examiners for Brief and Abbreviated IQs

To examine the inter-rater reliability IQs, one correlation was conducted between both examiners and the IQs from each test, resulting in a total of four correlations. Paired-samples *t*-tests were conducted in order to evaluate the mean differences in IQs between examiners for the IQs from each test. In instances of restriction or expansion of range in the IQs, the correlation coefficients were corrected using the same method as for the convergent validity. Table 5 includes uncorrected correlations and corrected correlations across examiners, mean differences across examiners, and the results of correlated-samples *t* tests for each test and subtest or composite score. Uncorrected inter-rater reliability coefficients ranged from .99 (KBIT-2 and WJ COG) to .83 (SB-V). After correcting for range restriction in scores, the inter-rater reliability corrected coefficients

ranged from 1.00 (WJ COG) to .94 (SB-V). Inter-rater reliability for IQ tests is considered adequate when $r > .80$. The uncorrected and corrected correlations for each test meet this standard, and the corrected correlations are strong to extremely strong. The WJ COG III and the KBIT-2 require the least amount of examiner judgment when scoring, and they produced the strongest correlations in IQs across examiners. Mean differences were approximately 1 standard score point or less, and all *t* tests revealed nonsignificant mean differences, $p > .05$.

Dependability Analysis

Finally, IQs were entered into a Generalizability theory analysis to examine their consistency and dependability. Variance components were computed using PASW 18.0, and dependability coefficients (a.k.a., phi coefficients) were calculated to provide overall indexes of dependability (Brennan, 2001; Shavelson & Webb, 1991). The variance estimate attributable to differences across the IQs was considered universe score variance; it was used as the numerator in the formula to calculate the dependability coefficients. The variance estimates attributable to the test, to examiners, to all interactions, and to residual (i.e., unexplained) variance, are then divided by the number of variations associated with each facet, resulting in error variance. The denominator of the formula consisted of the sum of the universe-score variance and error variance.

Table 7 provides the variance components estimates for examinee, examiner, test, and all interactions. For reference, the object to measurement, variance attributable to individual differences across examinees accounted for less than half of the variance; in fact, it accounted for only 41% of the variance. Thus, the remainder of variance was due to systematic or random error. When considering error variance components, the largest

proportion of variance was attributed to the test-by-examinee interaction; it accounted for 46% of the variance. These results support the hypothesis that the test-by-examinee interaction component would contribute the largest variance in IQs of any error variance component. The test contributed 7% of the variance, supporting the hypothesis that the test component would contribute small but notable variance in IQs. However, the examiner, examinee-by-examiner interaction, and examiner-by-test interaction did not contribute to the variance in scores. These results did not support the hypothesis that the examiner-by-test component would be sizeable. Residual variance was only 6%. The dependability coefficient for all components of exchangeability was .75, indicating suspect dependability.

Discussion

It is possible to efficiently examine the influences of test properties and the effect of examiners on the exchangeability of IQs using characteristics of brief and abbreviated intelligence. Frequently, correlations between IQs and mean differences across IQs are used by researchers to examine the effects of inter-rater reliability and convergent validity for different tests. However, due to limitations in administering full-length IQ tests, this study used brief and abbreviated tests. This study was different from most studies because it examined the effects of both the inter-rater reliability, which has only minimally been studied, and the convergent validity. It is important that both of these influences were examined in one study because it helped determine their total effect on exchangeability of brief and abbreviated IQs as well the strength of their effects.

Dependability and Exchangeability of Brief and Abbreviated IQs

When test variance, examiner variance, and their interactions were considered collectively, the resulting dependability coefficient was weaker than the internal consistency reliability coefficients for each test in isolation. For example, despite very strong mean reliability coefficients (i.e., .91 to .96), the dependability coefficient was .75, indicating suspect dependability when compared to the .80 requirement for a test to be considered reliable. The suspect dependability in this study is similar to what Floyd et al. (2008) found when examining IQs from three intelligence tests (dependability coefficient = .53). The remainder of the pairwise IQ comparisons across samples yielded higher dependability coefficients (M = .73) than the dependability coefficient for three tests. All of the coefficient values were below the minimal standards for reliability as well as the internal consistency values for each IQ. It is possible that controlling for maturation by administering all tests in a single session contribute to the higher dependability coefficient in the current study, which was not a focus of the Floyd et al. study.

The results from this study, which evaluated more than a single error variance component, are also similar to Bergeron, Floyd, McCormack, and Farmer (2008). They found suspect dependability while evaluating the exchangeability of behavior rating scales across three types of error variance components--time, rater, and instrument—and their interactions; resulting dependability coefficients ranged from .47 to .68.

Although the brief and abbreviated IQs targeted in this study demonstrated adequate reliability (e.g., internal consistency and test-retest reliability) and moderate correlations with each other when error sources were examined in isolation, the simultaneous effects of differing tests, differing examiners, and all interactions led to

weaken the dependability of these IQs. In fact, low dependability was due to the fact that error variance components contributed more variance to IQs than the object of measurement, individual differences in general intelligence.

Test effects. Similar to the results from Floyd et al. (2008), the test component contributed 7% of the variance in brief and abbreviated IQs. This slight variance may be attributable to the Flynn effect (Flynn, 2006), which was supported by comparisons of the means across the tests. The KBIT-2 and SB-V are the most recently published tests, and their IQs were 2-7 points lower than the WASI and WJ COG BIA. As hypothesized, the WASI (the test normed earliest) produced IQs that were significantly higher than those from the KBIT-2 and SB-V.

Test-by-examinee effects. Ideally, the largest variance would be due to the individual differences between the examinees. The results from this study indicate that only 41% of the variance is due to the examinees, whereas the largest variance component was the test-by-examinee interaction (46%). In short, some students performed well on some tests, whereas others performed better on other tests. There are several possible explanations for the larger variance contribution from the test by examinee variance, the most notable of which concerns the differences in item scaling and adequate floors and ceilings.

In addition, the content and presentation of items within subtests contributing to intelligence tests may vary substantially. It is possible that these variations and the examinee's interaction with them might result in different scores across tests (McGrew 2009a, 2009b). For example, some tests require verbal responses to items, whereas other intelligence tests require few verbal responses and rapid motor responses (e.g., Visual

Matching on the WJ III). As a result, a child with strong verbal abilities may score higher on a test with lots of verbal items and fewer rapid motor response items but lower on a test with lots of rapid motor response items and fewer verbal items.

Examiner effects. Results of the Generalizability theory analysis showed that the variance due to the examiner and the component representing interactions with examiner effects contributed negligible variance in IQs. In contrast to what was originally hypothesized, the tests requiring more subjectivity did not have significantly lower inter-rater reliability. Furthermore, the results from the inter-rater reliability analysis were congruent with the Generalizability theory analysis. For example, the corrected inter-rater reliability for each test ranged from .94 to 1.00 (M = .97), which is well above the .90 criteria. The Generalizability theory and inter-rater reliability analyses indicate that inter-rater reliability is not a major limitation (on a relative scale) because there was negligible variance due to examiner by test and examiner alone.

Error effects. The error variance for this study contributed only 6% of the difference between scores. As a result, there are few unaccounted for confounds that contributed to the differences in IQs. It is notable, and perhaps a coincidence, that about the same percentage of variance in each IQ (based on internal consistency reliability estimates) could be attributed to random error.

Limitations

Due to the nature of the sample, the results cannot be extended to the general population. College students are usually not administered IQ tests, rather the tests are usually administered within school settings in order to determine a child's eligibility for special education. Therefore, the results of this investigation may not be comparable for

other populations. Additionally, none of the participants reported being previously diagnosed with a learning or developmental disorder, which makes it difficult to extend results to individuals with learning problems and developmental delays. Furthermore, in order to be admitted to college the students were required to have a minimum ACT score of 16 that might have contributed to the restriction of range.

Although this study employed an innovative method to examine the test and examiner effects in a single study, the procedure of recording and later scoring responses to test items was not perfect and contributed error to scoring. Several responses were difficult to hear and a few were inaudible, making it difficult to accurately score some items. Also, some types of errors were administration errors that could not be corrected by the secondary examiner (e.g., failure to establish a basal or ceiling), which systematically decreases the accuracy of the IQs obtained by the secondary examiner. However, these errors likely had minimal effect on the IQs the secondary examiner obtained.

Implications

Ideally, the largest percentage of variance in the Generalizability theory analysis would be due to the individual differences across the examinees; however, the results from this study indicate that the largest variance component is the test-by-examinee interaction. For this reason, examiners must be careful in choosing a test to ensure that it has adequate floors and ceilings depending on the referral concern. In addition, it may be beneficial to administer 2 tests in order to assure that the obtained scores are indicative of the examinee's skill level.

The results from the Generalizability theory analysis indicate that about 7% of the variance in IQ scores is from the test. Thus, because of a small, but notable effect of the tests on the Generalizability theory analysis, examiners should choose the most up-to-date tests when conducting assessments in order to obtain the most accurate IQ for an individual.

The variance due to examiner and all interactions with the examiner resulted in negligible differences between IQs. Despite this minimal variance, test manuals should begin to include inter-rater reliability for the entire test instead of select subtests in order to help examiners choose the most appropriate test. Unfortunately, most of the research focuses on inter-scorer agreement instead of inter-rater reliability and has targeted scoring of only Verbal subtests from the Wechsler scales, which use a three-point scale (e.g., 0, 1, and 2 points) based on sample responses and general criteria (e.g., degree of abstraction) shown in the manuals. For this reason, more research should be conducted on inter-rater reliability for whole tests, including subtests that do not require subjectivity in scoring. Further research should be conducted to determine if these differences are similar for other populations including children and individuals with known learning problems.

References

Alfonso, V. C., Johnson, A., Patinella, L., & Rader, D. E. (1998). Common WISC-III examiner errors: Evidence from graduate students in training. *Psychology in the Schools, 35*, 119-125.

Barrett, P. (2002). Attenuation corrections (v2.1).

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

Bergeron, R., Floyd, R. G., McCormack, A. C., & Farmer, W. L. (2008). The generalizability of externalizing behavior composites and subscale scores across time, rater, and instrument. *School Psychology Review, 37*, 91-108.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 4*, 313-326.

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

Bradley-Johnson, S. (1987). *Cognitive Ability Scale.* Austin, TX: Pro-Ed.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Camara, W. J., Nathan, J. S., & Puente, A. S. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31,*141-154.

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

Elliott, C. D. (1990). *The Differential Ability Scales.* San Antonio, TX: The Psychological Corporation.

Erdodi, L. A., Richard, D. C., & Hopwood, C. (2009). The importance of relying on the manual: Scoring error variance in the WISC-IV Vocabulary subtest. *Journal of Psychoeducational Assessment, 27,* 374-385.

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice, 39,* 414-423.

Flynn. J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law, 12,* 170-189.

Flynn, J. R. (2009). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16*(2), 98-104.

Homack, S. R., & Reynolds, C. R. (2007). *Essentials of assessment with brief intelligence tests.* A. S. Kaufman & N. L. Kaufman (Eds.). New York, NY: Wiley.

Hunnicutt, L. C., Slate, J. R., Gamble, C., & Wheeler, M. S. (1990). Examiner errors on the Kaufman Assessment Battery for Children: A preliminary investigation. *Journal of School Psychology, 28,* 271-278.

Jensen, A. R. (1993). Spearman's Generalizability factor: Links between psychometrics and biology. In F. M. Crinella & J. Yu (Eds.), *Brain mechanisms: Papers in memory of Robert Thompson* (pp. 103-129). New York: New York Academy of Sciences.

Kaufman, A. S., Kaufman, J. C., Balgopal, R., & McLean, J. E. (1996). Comparison of three WISC-III short forms: Weighing psychometric, clinical, and practical factors. *Journal of Clinical Child Psychology, 25,* 97-105.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test, Second Edition: Manual.* Circle Pines, MN: American Guidance Service.

Kuentzel, J., Hetterscheidt, L., & Barnett, D. (2011). Testing intelligently includes double-checking Wechsler IQ scores. *Journal of Psychoeducational Assessment, 29,* 39-46.

McGrew, K. S. (2009a).The Standard error of measurement (SEM): An explanation and facts for "Fact Finders" in Atkins MR/ID death penalty proceedings. *Applied psychometrics 101: IQ test score difference series.* Retrieved from http://www.iapsych.com/iapap101/iapap101_5.pdf

McGrew, K. S. (2009b). Understanding global IQ test correlations. *Applied psychometrics 101: IQ test score difference series.* Retrieved from http://www.iapsych.com/iapap101/iapap101_1.pdf

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities: Technical manual.* Itasca, IL: Riverside Publishing.

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

Moon, G. W., Blakey, W. A., Gorsuch, R. L., & Fantuzzo, J. W. (1991). Frequent WAIS-R administration errors: An ignored source of inaccurate measurement. *Professional Psychology: Research and Practice, 22*, 256-258.

Peterson, D., Steger, H. S., Slate, J. R., Jones, C. H., & Coulter, C. (1991). Examiner errors on the WRAT-R. *Psychology in the Schools, 28*, 205-208.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*, 33-65.

Ramos, E., Alfonso, V. C., & Schermerhorn, S. M. (2009). Graduate students' administration and scoring errors on the Woodcock-Johnson III Tests of Cognitive Abilities. *Psychology in the Schools, 46*, 650-657.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition, Technical Manual.* Itasca, IL: Riverside Publishing.

Ryan, J. J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R. *Journal of Consulting and Clinical Psychology, 51*, 149-150.

Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III). *Assessment, 10*, 151-159.

Ryba, N. L., Cooper, V. G., & Zapf, P. A. (2003). Juvenile competence to stand trial evaluations: A survey of current practices and test usage among psychologists. *Professional Psychology: Research and Practice, 34***,** 499-507.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools, 26,* 78-84.

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

Slate, J. R., & Jones, C. H. (1990a). Examiner errors on the WAIS-R: A source of concern. *The Journal of Psychology, 124,* 343-345.

Slate, J. R., & Jones, C. H. (1990b). Identifying students' errors in administering the WAIS-R. *Psychology in the Schools, 27,* 85-87.

Slate, J. R., Jones, C. H., Coulter, C., & Covert, T. L. (1992). Practitioners' administration and scoring of the WISC-R: Evidence that we do err. *Journal of School Psychology, 30,* 77-82.

Tabachnick, B. G., & Fidell, L. S. (2006) *Using multivariate statistics, 5th ed.* Boston: Allyn and Bacon.

Terman, L. M., & Merrill, M. A. (1937). *Measuring intelligence.* Boston: Houghton Mifflin.

Ward, L. C., & Ryan, J. J. (1996). Validity and Time Savings in the Selection of Short Forms of the Wechsler Adult Intelligence Scale—Revised. *Psychological Assessment, 8,* 69-72.

Wechsler, D. (1974). Wechsler Intelligence Scale for Children, Revised Edition. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1981). Wechsler Adult Intelligence Scale, Revised Edition. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1989). Wechsler Preschool and Primary Scale of Intelligence, Revised Edition. San Antonio, TX: The Psychological Corporation.

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

Wechsler, D. (1997). Wechsler Adult Intelligence Scale, Third Edition. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1999). *Wechsler Abbreviated Scales of Intelligence Manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). Wechsler Intelligence Scale for Children, Fourth Edition. San Antonio, TX: The Psychological Corporation.

Wilson, M. S., & Reschly. D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review.* 25. 9-23.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2006). *Woodcock Johnson III Normative Update Compuscore and Profiles*. Itasca, IL: Riverside Publishing.

Table 1
Summary of Research Examining Examiner Errors

Year	Authors	Examiners	Test	Design/Trial	Results
1989	Slate & Chick	14 graduate students enrolled in a clinical psychology program	WISC-R	Each student administered the WISC-R 8 times to child or adolescent volunteers	Vocabulary, Similarities, and Comprehension were the three subtests that contained the highest number of errors. The average number of errors per protocol was 15.2. Examiners were more likely to give more credit for responses than the manual stated resulting in inflated FSIQs. The FSIQ was unchanged due to error in only 32.6% of the protocols. The corrected IQs were generally 1 to 3 points lower than the corrected score.
1990	Hunnicutt et al.	Licensed clinical and school psychologists	KABC	Investigated errors on 46 protocols	Found that 83% contained at least one error and about half of them contained errors that changed the overall IQ.
1990a	Slate & Jones	22 master's level clinical psychology graduate students	WAIS-R	Each participant administered 7 practice submissions of the WAIS-R	After completion of the 7 practice submissions there were 149 protocols to evaluate (1 participant only turned in 5 protocols and another only turned in 4). Based on these protocols the number of errors committed on each protocol. There were 145 protocols with errors in them resulting in either inflated or deflated IQs.

Table 1

Summary of Research Examining Examiner Errors

Year	Authors	Examiners	Test	Design/Trial	Results
1990b	Slate & Jones	26 graduate students	WAIS-R	26 students were randomly chosen to administer the WAIS-R 5 to 8 times	180 WAIS-R protocols from 26 graduate students were found to have an average of 8.8 errors per protocol. Correction of errors revealed that 81% of the FSIQ scores changed
1991	Moon et al.	33 school psychology doctoral students:	WAIS-R	Each participant completed the Criteria for Competent WAIS-R Administration (CCWA)	The CCWA was completed in order to determine specific errors of omission or commission and an overall accuracy of administration. This study also looked at the most common errors that graduate students committed across two administrations of the WAIS-R.
1991	Peterson, Steger, Slate, Jones, & Coulter	9 school psychologists	WRAT-R	A random sample of 55 protocols from 9 examiners, each examiner completed an average of 6.11 protocols (range 1-12).	All nine psychologists made errors with an average of 3 errors per protocol. Some of the most common errors were inaccurate basals and ceilings and failure to record responses.

Table 1

Summary of Research Examining Examiner Errors

Year	Authors	Examiners	Test	Design/Trial	Results
1992	Slate, Jones, Coulter, & Covert	8 licensed psychological examiners 1 certified educational examiner	WAIS-R	56 randomly selected psychological folders from a school system were evaluated	56 WISC-R protocols from 9 certified examiners were found to have an average of 38.4 errors per protocol. Errors on 81% of the protocols resulted in changes in the FSIQ score
1998	Alfonso, Johnson, Patinella, & Rader	15 graduate students	WISC-III	Each participant administered 4 WISC-III for training	The most common errors were studied. The most common errors include: failure to query or record answers verbatim, reporting FSIQ incorrectly, reporting VIQ incorrectly, and incorrectly adding raw scores.
2009	Erdodi, Richard, & Hopwood	46 clinical psychology graduate students	WISC-IV	Each participant scored 3 partially completed Vocabulary subtests from the WISC-IV	There were three protocols each one was designed to produce a different scaled score (4, 10, and 16). The results showed that raters were more likely to award too many points and most errors occurred on the protocols for the extreme scores. For the lowest scaled score 75% produced a higher scaled score. For the highest scaled score 67% produced a higher scaled score.

Table 2
Summary of Research Examining Inter-Rater Reliability

Year	Author(s)	Examiners	Intelligence Test	Design/Trials	Results
1980	Bradley, Hanna, & Lucas	63 members of NASP	WISC-R	Each participant scored 2 WISC-R protocols	Each protocol was for a 10-year-old female, one was explicitly easy to score and the other was designed to be more difficult to score. The results showed that the FSIQ can vary by as much as 6-8 points depending on the examiner.
1983	Ryan et al.	19 school psychologists 20 school psychology graduate students	WAIS-R	Each participant scored 2 WAIS-R protocols for the same 2 clients from a vocational psychology clinic	Results showed that errors made in scoring caused IQs to vary from as much as 4 to 18 points, regardless of level of training.
2003	Ryan & Schnakenberg-Ott	19 School psychologists PhD. 19 school psychology graduate students	WAIS-III	Each participant scored 2 WAIS-III for the same 2 clients from a neuropsychology clinic	Evaluated how often an obtained IQ fell outside of the +/-4 confidence interval of the actual obtained IQ and if the obtained IQ fell into a different ability range. This study showed that regardless of one's level of training that errors can be made; therefore detracting from the accuracy of the obtained IQ.

Table 3

Comparisons of Brief Tests with other Brief and Full-Scale Tests

Brief Intelligence Tests	*Comparison Battery*	*Score*	*Participants*	*Correlation*
Kaufman Brief Intelligence Test, Second Edition (KBIT-2; Kaufman & Kaufman, 2004)	KBIT	IQ Composite	Adults ages 16-45	.82
	WASI	FSIQ-4	Adults ages 35-52	.90
	WASI	FSIQ-4	Children ages 7-19	.76
	WASI	FSIQ-2	Adults ages 35-52	.88
	WASI	FSIQ-2	Children ages 7-19	.71
	WISC-III	FSIQ	Children ages 6-15	.78
	WISC-IV	FSIQ	Children ages 6-16	.66
	WAIS-III	FSIQ	Adults ages 20-48	.89
Wechsler Abbreviated Scales of Intelligence (WASI; Wechsler, 1999)	WISC-III	FSIQ-2	Children ages 6-16	.81
	WAIS-III	FSIQ-2	Adults ages 16-89	.87
Woodcock Johnson III Tests of Cognitive Abilities (WJ III COG; McGrew & Woodcock, 2001) Brief Intellectual Ability (BIA)	DAS	GCA	Children ages 1-6	.67
	WPPSI	FSIQ	Children ages 1-6	.67
	SB-V	FSIQ	Age range of 3-5	.60
	WISC-III	FSIQ	Children ages 8-12	.69
	WJ III COG	GIA	Norm sample	.92
	DAS	GCA	Children ages 8-12	.70
	CAS	FSS	Children ages 5-14	.70
	WAIS-III	FSIQ	College students ages 18-53	.62
Stanford-Binet, Fifth Edition (SB-V; Roid, 2003) Abbreviated Battery IQ (ABIQ)	SB-V	FSIQ	Ages 6+	.87

Note. KBIT = Kaufman Brief Intelligence Test; WASI = Wechsler Abbreviated Scales of Intelligence; FSIQ = Full Scale IQ; WISC-III = Wechsler Intelligence Scales for Children, Third Edition; WISC-IV = Wechsler Intelligence Scales for Children, Fourth Edition; WAIS-III = Wechsler Adult Intelligence Scales, Third Edition; DAS = Differential Abilities Scales; WPPSI = Wechsler Preschool and Primary Scales of Intelligence; SB-V = Stanford Binet, Fifth Edition; WJ III COG = Woodcock Johnson III Tests of Cognitive Abilities; CAS = Cognitive Assessment System.

Table 4

Secondary Examiner Corrections using Random Numbers List

Type of Error	WASI	KBIT-2	SB-V	WJ III COG
Failure to Query examinee response (e.g., verbal responses listed in the manual)	68	0	1	9
Inaudible responses (e.g., unable to hear what the examinee said)	0	2	0	1
Mispronunciations of vocabulary words	1	3	32	0
Failure to Prompt (e.g., prompt for one-word responses, prompt for narrowing response)	0	0	0	30
Didn't Repeat Item ?	0	1	0	0
No Corrective Feedback (e.g., on teaching or sample items)	0	1	0	1
Total Errors	69	7	33	41

Table 5
Means, Standard Deviations, and Inter-Rater Reliability Correlations for IQs

	Primary examiner			Secondary examiner			Inter-rater reliability			
IQs	*M*	*SD*	Range	*M*	*SD*	Range	*r*	r_c	*M* diff*	*t*
KBIT-2 Composite	97.03	11.97	74-117	97.13	11.37	75-117	.99	.99	-.10	-.32
WASI FSIQ-2	103.10	10.91	81-123	103.59	11.05	80-127	.90	.95	-.49	-.62
WJ III COG BIA	99.18	9.89	81-124	99.15	10.82	76-128	.99	1.0	.03	.08
SB-V ABIQ	96.10	9.03	73-112	97.15	8.37	76-115	.83	.94	-1.05	-1.29

Note. Composite and subtest scores are age-based standard scores (*M* = 100, *SD* = 15) unless otherwise noted. KBIT-2 = Kaufman Brief Intelligence Test, Second Edition; WASI = Wechsler Abbreviated Scales of Intelligence; WJ III COG = Woodcock Johnson III Tests of Cognitive Abilities; BIA = Brief Intellectual Ability; SB-V = Stanford Binet, Fifth Edition; ABIQ = Abbreviated IQ.
All correlations are statistically significant, $p < .001$ (two-tailed).
**M* diff = Mean differences between examiners.

Table 6
Correlation matrix for tests by examiner

	Obtained correlations			
Measure	1	2	3	4
1. KBIT-2 IQ Composite	-	.76**	.59**	.18
2. WASI FSIQ-2	.76**	-	.38*	.32*
3. WJ III COG BIA	.63**	.48**	-	.17
4. SB-V ABIQ	.29	.26	.15	-
	Corrected correlations			
Measure	1	2	3	4
1. KBIT-2 IQ Composite	-	.88	.79	.62
2. WASI FSIQ-2	.88	-	.69	.68
3. WJ III COG BIA	.82	.74	-	.60
4. SB-V ABIQ	.67	.63	.57	-

Note. Pearson product-moment correlation coefficients for the Primary Examiner are presented below the diagonal, and correlations for the Secondary Examiner are reported above the diagonal. We also recognize that there is no set standard for providing nominal labels for *r* values. KBIT-2 = Kaufman Brief Intelligence Test, Second Edition; WASI FSIQ-2 = Wechsler Abbreviated Scale of Intelligence, Full-Scale IQ; WJ III COG BIA = Woodcock Johnson III, Tests of Cognitive Abilities, Brief Intellectual Abilities; SB-V ABIQ = Stanford Binet, Fifth Edition, Abbreviated IQ.
*$p < .05$ (two-tailed).
**$p < .001$ (two-tailed).

Table 7
Variance Component Estimates and Absolute Dependability Coefficients by Score Comparison

	Estimated variance components	
Facet	Brief or abbreviated IQs	Percent of variance
Examinee	50.06	41%
Examiner	0.05	0%
Test	8.39	7%
Examinee-by-examiner	0.11	0%
Examinee-by-test	55.43	46%
Examiner-by-test	-0.07 [a]	0%
Residual	7.08	6%
Total	121.12	
ϕ	.75	

Note. [a] = Negative estimated variance components were set to zero.

Appendix A

Participant Demographics Form

Tell Us About <u>Yourself</u> by Completing the Blanks or by Placing Checks in the Boxes

Your Date of Birth:_____

Your College Classification:
Freshman Sophomore
Junior Senior

Your Gender: Male Female

Your Race (please check only one):
African American/Black White/Caucasian Asian/Pacific Islander
Native American/American Indian Arab American Biracial or Multiracial
Other (please specify)_____

Are you of Hispanic origin? Yes No
If yes, what is your family's country of origin?_____

Have you been diagnosed with ADHD or another behavior, emotional, or learning problem? Yes No
If yes, which diagnoses?
_____

Do you take a prescription medication on a regular basis? Yes No
If yes, what medication? _____

What is your current occupation? _____

Please check the highest grade level or degree completed by each of your **<u>parents or caregivers</u>**:

Mother/Female Caregiver (if applicable)	**Father/Male Caregiver** (if applicable)
Less than High School Diploma or GED	Less than High School Diploma or GED
High School Diploma or GED	High School Diploma or GED
Some College	Some College
Technical School	Technical School
Bachelor's Degree	Bachelor's Degree

Higher than Bachelor's Degree Higher than Bachelor's Degree
Current occupation:_____ Current occupation:_____

From what University of Memphis course were you recruited or will you received credit? Provide name._____

Appendix B

Examiners Demographic Form

1. Gender (circle one): Male Female
2. Ethnicity: _____
3. Program (circle one): MS/PhD MA/EdS
4. Number of graduate hours completed: _____
5. Approximate number of practicum hours completed
6. Number of tests administered (including practice):
    a. Wechsler (WISCs, WAISs, WASIs): _____
    b. WJ III COG: _____
    c. RIAS: _____
    d. WJ III ACH _____
    e. WIAT-II _____
    f. Early Numeracy _____
    g. Early Literacy _____
    h. Other: _____
7. Number of psychoeducational assessment reports written in practice:

    _____

THE UNIVERSITY OF MEMPHIS

Institutional Review Board

To: Sarah Irby
Psychology

From: Chair, Institutional Review Board
For the Protection of Human Subjects
irb@memphis.edu

Subject: Exchangeability of Brief IQs (091610-73)

Approval Date: November 5, 2010

This is to notify you of the board approval of the above referenced protocol. This project was reviewed in accordance with all applicable statuses and regulations as well as ethical principles.

Approval of this project is given with the following obligations:

1. At the end of one year from the approval date, an approved renewal must be in effect to continue the project. If approval is not obtained, the human consent form is no longer valid and accrual of new subjects must stop.
2. When the project is finished or terminated, the attached form must be completed and sent to the board.
3. No change may be made in the approved protocol without board approval, except where necessary to eliminate apparent immediate hazards or threats to subjects. Such changes must be reported promptly to the board to obtain approval.
4. The stamped, approved human subjects consent form must be used. Photocopies of the form may be made.

This approval expires one year from the date above, and must be renewed prior to that date if the study is ongoing.

_____

Chair, Institutional Review Board
The University of Memphis

Cc: Dr. R. Floyd