4-17-2014

The Exchangeability of Brief Intelligence Tests for Children with Intellectual Giftedness: Illuminating Error Variance Components' Influence on IQs

Sarah McCallum Irby

THE EXCHANGEABILITY OF BRIEF INTELLIGENCE TESTS FOR CHILDREN WITH INTELLECTUAL GIFTEDNESS: ILLUMINATING ERROR VARIANCE COMPONENTS' INFLUENCE ON IQS

by

Sarah McCallum Irby

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Psychology

The University of Memphis

May 2014

Abstract

Irby, Sarah McCallum. Ph.D. The University of Memphis. May 2014. The exchangeability of brief intelligence tests for children with intellectual giftedness: Illuminating error variance components' influence on IQs. Major Professor: Randy G. Floyd.

This study examined the exchangeability of IQs from three brief intelligence tests. Tests were administered to 36 children with intellectual giftedness, scored by one set of primary examiners and later scored by a secondary examiner. For each student, 6 IQs were calculated and submitted to a Generalizability theory analysis. Despite strong convergent validity and reliability evidence supporting brief IQs, the resulting dependability coefficient was only .80, which indicates relatively low exchangeability across tests and examiners. Although error variance components representing the effects of the examiner, examiner-by-examinee interaction, the examiner-by-test interaction, and the test contributed little to IQ variability, the component representing the test-by-examinee interaction contributed about one-third of the variance in IQs. These findings hold implications for selecting and interpreting brief intelligence tests and general testing for intellectual giftedness.

Keywords: intellectual giftedness, IQ, Generalizability theory, dependability coefficients

Table of Contents

Exchangeability of Brief Intelligence Tests: Illuminating Error Variance Components' Influence on IQs for Children with Intellectual Giftedness

Ensuring accuracy of measurement in psychology, as well as other social sciences, the physical sciences, and sports is of vital importance. Measurement accuracy is of utmost importance when lives are at stake. For example, in the physical sciences, knowing the elevation of an airplane is essential for landing safely and different altimeters (e.g., using barometric pressure or radar) and different people using the same method should ideally produce similar results (i.e., landing the airplane safely). However, an altimeter reading may be more or less accurate depending on the speed of the plane or if the pilot is tired. Although, the accuracy of the altimeter is important at all levels, it is perhaps more important during the landing, where the difference of a few feet could mean the difference between the life and death of passengers. Measurement is also important in sports and leisure. For example, timing of track and field and recreational races is central to these sports and different methods of timing (e.g., timing by judges, self-timing, and timing using electronic methods producing a “chip time”) and different people using the same method should ideally produce similar times. However, timing of track and field and recreational races may be more or less accurate depending on the speed at which participants run. Although timing accuracy should be ensured at all levels, it is perhaps most important that timing be correct for those running the fastest. Awards, medals, and race-specific, regional, national, and international records are determined based on the timing of these runners. Precise measurements of altitude during landing and time for the most elite runners can be considered high-stakes, whereas altitude during flight and timing of runners lagging behind appears to be less relevant. This same pattern is evident in educational and psychological measurement—especially when assessing children and adolescents with intellectual giftedness.

Intellectual giftedness.

According to the United States Department of Education (2000), approximately 6-7% of students enrolled in public elementary schools meet the criteria for intellectual giftedness. Section 806 of Public Law 91-230 defines intellectual giftedness as an individual who has high capability in one or more of the following areas: (a) cognitive functioning, (b) creative thinking, (c) leadership skills, (d) visual and performing arts, and (e) specific ability areas. This federal law was implemented in 1971, but the specifics of eligibility criteria and funding were left to the state and local governments, and differentiated education for children who are intellectually gifted has been a low priority in regards to funding at the federal, state, and local levels (Marland, 1972). Furthermore, Stephens (2012) argued that more recent education policies have not been directly focused on identifying and nurturing the academic potential of children with intellectual giftedness. Instead, these policies have focused on the achievement of proficiency or minimum competency and providing resources to underperforming students.

McClain and Pfeiffer (2012) surveyed state gifted consultants from each state and asked about eligibility criteria and cut-scores for gifted eligibility. A cut-score is a single point on an IQ or achievement score continuum that differentiates between one condition and another. In the case of giftedness, the differentiation is between intellectual giftedness and average intelligence. McClain and Pfeiffer found that each state uses different criteria for eligibility including what they called (a) a single cutoff–flexible criterion, (b) multiple cutoffs, and (c) no model. Seven states (14%) utilize the single cutoff–flexible criterion, which uses one piece of diagnostic information as determined by the local school district. For example, a school district may use a cut-score of 130 on an intelligence test, but another district may use a cut-score of 125 on an achievement test. Twenty-seven (54%) states use multiple cutoffs, which require scores at or above a specific cut-score on multiple measures (e.g., intelligence test, creativity test, or state achievement test). Additionally, McClain and Pfeiffer found that 15 states (36%) require a specific intelligence test cut-score for gifted eligibility (i.e., 120, 125, or 130). Furthermore, of these 15 states, 7 states (47%) require a cut-score above 130, 7 states (47%) require a cut-score above 125, and 1 state (6%) requires a cut-score above 120.

A comprehensive assessment to determine eligibility for intellectual giftedness usually includes a review of records including the most recent state standardized achievement test scores (e.g., the Tennessee Comprehensive Assessment Program or TCAP). Additionally, the assessment will include an intelligence test and will likely include an assessment of creativity (e.g., Torrance Tests of Creative Thinking; Torrance, 1974). Ultimately, the majority of states include an intelligence test as part of a comprehensive assessment for intellectual giftedness, regardless of how much weight is placed on the IQ alone. However, students are usually identified as intellectually gifted based on individual scores on norm-referenced intelligence tests (Robinson, 2005; Worrell, 2009).

An intelligence test is designed to measure an individual’s cognitive abilities; general intelligence is a common source of individual differences found in assortment of cognitive tasks (Jensen, 1998). However, each test goes about measuring general intelligence in a slightly different way. These differences can be problematic in gifted eligibility assessments 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.

High-stakes assessments are required for determining if a child is intellectually gifted; therefore, it is important to be mindful of several important issues related to IQs before accepting a score as valid and determining eligibility for intellectual giftedness based, in part, on only one IQ from a single test. Moreover, it is important to understand the *exchangeability* of IQs for children with intellectual giftedness. Exchangeability refers to the likelihood that IQs are the same despite the varying conditions under which they are obtained (Floyd, Clark, & Shadish, 2008). For example, a child's performance on one intelligence test may lead them to meet eligibility criteria for intellectual giftedness, whereas the child's performance on another intelligence test may not. Additionally, examiner error may also affect a child's score on an intelligence test, resulting in a possible underidentification or overidentification of gifted students. Exchangeability can be investigated using a variety of group-based and person-centered analyses, including mean differences, convergent validity, inter-rater reliability analysis, and score confidence interval overlap. Another way to investigate exchangeability is through a Generalizability theory (G-theory; Shavelson & Webb, 1991) analysis. The G-theory, in addition to the overall estimate of dependability, can also produce estimates of the sources of true score and error variance. Variance from test characteristics, examiner effects, and all interactions should be considered in concert with a population of children with intellectual giftedness in order to understand the exchangeability of IQs during high-stakes assessments.

IQ Exchangeability and Sources of Error Variance

Convergent relations between IQs from varying intelligence tests have been examined across hundreds of studies, but their exchangeability has only recently been targeted (Floyd et al., 2008; Irby & Floyd, 2011; Irby, Floyd, & Bergeron, 2013). Previous research examining the exchangeability of 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 strong correlation of .84 between the General Conceptual Ability (GCA) from the Differential Abilities Scale, Second Edition (DAS-II; Elliott, 2007) and the Full-Scale IQ from the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Wechsler, 2003). However, correlations are not sensitive to absolute score differences, so in order to better understand score exchangeability, it is also necessary to evaluate mean differences in IQs from two or more different tests. Moreover, these analyses (e.g., correlations and mean differences) provide information for only a pair of IQs. In order to evaluate the exchangeability of multiple IQs simultaneously, a G-theory analysis can be conducted. The G-theory extends traditional notions of reliability measurement by partitioning variance in observed IQs. Thus, a G-theory analysis is useful in determining the variance components that may contribute to low exchangeability of IQs by separating the object of measurement (i.e., individual differences in general intelligence) from error variance components (e.g., test characteristics, examiner effects, and all interactions) and residual error (i.e., test-by-examiner-by-examinee effects). Furthermore, a dependability coefficient is yielded to help determine the accuracy of generalizing one person’s observed IQ on one intelligence test to the average score that person would have received under all possible circumstances (e.g., across tests and examiners; Shavelson & Webb, 1991).

Floyd et al. (2008) investigated several methods of studying exchangeability (e.g., Pearson *r*, mean differences of the absolute score, and the difference between the participants' pair of IQs) across seven intelligence tests and six 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. An example of these results includes the finding that more than 33% of participants who completed the Woodcock–Johnson III Tests of Cognitive Abilities (WJ III; McGrew & Woodcock, 2001) and Wechsler Intelligence Scale for Children, Third Edition (WISC-III; Wechsler, 1991) demonstrated a difference between IQs of more than 10 points. In addition, Floyd et al. completed G-theory analysis, which provided an estimate of the exchangeability of the IQ (the dependability coefficient) based on the variance due to individual differences in general intelligence compared to all other variance components. For example, an analysis employing IQs from the WISC-III and the WJ III from a child sample produced a dependability coefficient of .68, which did not meet the standards set by Nunnally and Bernstein (1994). They argue that basic research should require reliability coefficients to exceed .80, but when important decisions are being made (e.g., assessment of intellectual giftedness), a reliability coefficient of .90 is the minimum, with .95 the desirable standard (Nunnally & Bernstein, 1994). In addition, the remainder of the pairwise IQ comparisons across their 10 analyses using data from children, adolescents, and adults yielded somewhat stronger dependability coefficients (*M* = .73) than the previously mentioned dependability coefficient for three tests; however these coefficient values were still well below the internal consistency reliability values for each IQ and Nunnally and Bernstein’s (1994) recommendation. Despite this body of research, relatively little is known about the exact reasons for the relatively IQ exchangeability or the relative contributions of varying sources of error in producing IQ differences. It is thought that characteristics of the test (Floyd et al., 2008; Irby & Floyd, 2011; Irby et al., 2013) as well as differences due to the effects of examiners (see, for example, Ryan & Schnakenberg-Ott, 2003) and their interactions (i.e., test-by-examinee, test-by-examiner, and examiner-by-examinee) are the most powerful construct-irrelevant influences on IQs.

Test characteristics. Test characteristics are believed (Floyd et al., 2008; Irby & Floyd, 2011; Irby & Floyd, 2011; Irby et al., 2013) to have a strong influence on IQs. One test characteristic thought to influence score exchangeability is 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 over time (Flynn, 2006, 2009). 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, 2009). For example, the DAS-II produced higher IQs on average than the WJ III (*MD* = 2.56), which was normed approximately 6 years earlier (Elliott, 2007).

A second test characteristic that is believed to influence score 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 (e.g., children with intellectual giftedness; 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. Another potential influence on IQ exchangeability is the "regression toward the mean" phenomena, which means an individual with an extreme IQ (e.g., 130) is likely to perform closer to the mean on subsequent tests. For example, if a child is administered two or more intelligence tests, it is unlikely that they will obtain IQs above 130 on all tests. Overall, there appears to be several test characteristics that are likely to have an effect on IQ variability and potentially result in suspect dependability.

Examiner effects. In addition to test characteristics, examiner effects are also believed (Floyd et al., 2008; Irby & Floyd, 2011; Irby et al., 2013) to influence IQs. 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. Moreover, most studies have explored examiner effects (i.e., inter-scorer agreement or inter-rater reliability) in isolation (e.g., Alfonso, Johnson, Patinella, & Rader, 1998; Erdodi, Richard, & Hopwood, 2009; Ryan & Schnakenberg-Ott, 2003).

The majority of published studies that examine inter-scorer agreement of IQs have focused on scoring of only the Verbal subtests (i.e., Vocabulary and Similarities) from the Wechsler scales, which use a three-point scale (e.g., 0, 1, and 2 points) for scoring items based on sample responses and general criteria (e.g., degree of abstraction) shown in the manuals. Several 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, Hanna, & Lucas, 1980; Ryan, Prifitera, & Powers, 1983; Ryan & Schnakenberg-Ott, 2003). Due to examiner errors on the Verbal subtests, these studies indicated that there is only a 26% to 35% overall agreement in IQs. Moreover, because most current research on inter-scorer agreement (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.

Error variance components from G-theory. Relatively recent G-theory (Shavelson & Webb, 1991) studies have examined these sources of error variance (e.g., test characteristics and examiner effects) and their total effect on exchangeability. A G-theory analysis produces dependability coefficients, as well as estimates of the sources of the object of measurement (i.e., individual differences in general intelligence) and error variance (i.e., test characteristics, examiner effects, and all interactions—test-by-examinee, test-by-examiner, and examiner-by-examinee). For example, Floyd et al. (2008) used G-theory to examine the magnitude of the effects of general characteristics of intelligence tests on variability in IQs. For most comparisons, Floyd et al. found that the variance component associated with differences in test characteristics was negligible, contributing less than 4% of the variance for five of the six samples. However, in one sample of adults targeting IQs from the Wechsler Adult Intelligence Scales, Third Edition (WAIS-III; Wechsler 1997), the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993), and WJ III, the test contributed 22% of the total variance. The influence of the interaction between individuals and the test characteristics—and residual error—contributed sizable variance for all six samples. In fact, this variance component accounted for 7% to 27% of all variance in IQs across these samples. Thus, the systematic variance in IQs that is not due to individual differences in ability does not typically come from the test characteristics themselves; instead, it in part comes from individuals' responses to concrete aspects of the tests, such as test stimuli, task requirements, response requirements, or through some subtle effects associated with variation in the representativeness of normative samples at varying ages.

Despite producing 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. (2008) study demonstrated some weaknesses in evaluating the reasons for IQ exchangeability. First, it was limited in that it examined only one class of error variance components—those associated with the tests under study. As a result of this limitation, the interaction between the individual and the test could not be separated from residual error. From a conceptual perspective, it is necessary for all potential influences on IQ exchangeability (e.g., test characteristics, examiner effects, and their interactions—test-by-examinee, test-by-examiner, examiner-by-examinee) to be examined at once in order to better understand of the reasons for IQ differences during high-stakes assessment. Second, the results from Floyd et al. stemmed from archival data from studies conducted during the validation process for commercially available intelligence tests, and several potential confounds, such as carryover effects and the effects of delays between testing sessions, do not appear to have been carefully controlled during data collection or follow-up analysis.

As an extension of the study conducted by Floyd et al., Irby and colleagues (2011, 2013) evaluated the effects of several error variance components (i.e., test characteristics, examiner effects, and their interactions) in producing variation of IQs in relation to individual differences in general intelligence. Four brief intelligence tests were (a) administered to a sample of college students who did not have previous clinical diagnoses (e.g., learning disabilities or intellectual giftedness) and (b) scored by a set of primary examiners (i.e., advanced school psychology graduate students). Irby and colleagues used brief intelligence tests as proxies for full-length intelligence tests to control for potential confounds associated with the effects of delays between testing sessions; all tests were administered in counterbalanced order within a single testing session. Furthermore, using a novel method for evaluating examiner effects, each testing session was video recorded and scored by a secondary examiner.

Irby and Floyd (2011) investigated several methods of studying exchangeability (e.g., convergent validity, mean differences, and inter-rater reliability) across four brief tests and two examiners. However, the focal point of this study was the G-theory analysis (Irby & Floyd, 2011; Irby et al., 2013), in which error variance components representing examiner effects, test characteristics, the examiner-by-examinee interaction, the examiner-by-test interaction, and the test-by-examinee interaction were specified. Variance components representing examiner effects, test characteristics, the examiner-by-examinee interaction, and the examiner-by-test interaction contributed minimal variability in IQs. In contrast, the component representing the test-by-examinee interaction contributed about half of the variation in IQs (48%), and this component was greater than the object of measurement (i.e., individual differences in general intelligence; 39%). Despite strong reliability and convergent validity evidence supporting the IQs generated from the brief intelligence tests, the resulting dependability coefficient was .75, indicating that the IQs from the brief intelligence tests have inadequate accuracy in measurement of general intelligence that is due primarily to the test-by-examinee interaction. In short, some students performed well on some tests, whereas others performed better on other tests, in a manner independent of their level of general intelligence.

Purpose of the Study

The exchangeability for IQs is far more important for individuals with abilities at the upper and lower extremes of the normal curve, such as in contexts in which children are assessed for intellectual giftedness. In a school setting, a few IQ points may mean the difference between (a) special education eligibility, which may include intensive interventions and transition services provided in school settings and (b) continued restriction to the general education curriculum with minimal support. However, previous research has mostly evaluated exchangeability for samples of children and young adults without identified clinical and educational conditions, and no exchangeability studies have been completed with those expected to be uniformly at the extremes of the normal curve (e.g., intellectually gifted). Furthermore, despite evidence that both test characteristics and examiners scoring differences may have substantial effects on IQ exchangeability, most previous research has evaluated only one of these influences at a time and Irby and colleagues (2011, 2013) have only recently examined both convergently.

To examine IQ exchangeability with children at the extreme of the IQ distribution and to address both the effects of test characteristics and examiner effects, this study compared IQs from three brief intelligence tests scored by independent examiners using samples of children with intellectual giftedness. Thus, all brief intelligence tests were administered and scored by one set of primary examiners and also scored independently by a single secondary examiner—producing six IQs for each child. This study answered three questions:

1. First, what are the relations and mean differences in IQs across tests and IQs produced by different examiners? Based on Irby and colleagues (2011, 2013), it is hypothesized that there would be moderate to strong relations across tests and very strong relations across examiners.
2. Second, what amount of error variance is attributed to the test, the examiner, the test-by-examiner interaction, the test-by-examinee interaction, and the examiner-by-examinee interaction? Additionally, what is the overall dependability of IQs? It is hypothesized that several error variance components are likely to be larger for the sample of children with intellectual giftedness than for a sample of average children or adults (cf. Floyd et al., 2008; Irby & Floyd, 2011; Irby et al., 2013). Specifically, it is likely that the test characteristics would contribute about 5% to 10% of the variance and may be higher in the gifted sample than in previous research due to variation in the ceilings for subtests contributing to the IQs, which are less likely to affect variation in IQs for children of average intellectual functioning. Furthermore, based on results from previous research (e.g., Floyd et al., 2008; Irby & Floyd, 2011; Irby et al., 2013) and the Flynn effect (Flynn, 2006), it is hypothesized that more recent intelligence tests would yield lower IQs than those with earlier publication dates, which indicates that the test component would contribute to small but notable variance in IQs. Additionally, it is hypothesized that the overall dependability would be moderate to strong and

would likely suggest suspect reliability in IQs, based on Nunnally and Bernstein (1994).

Additionally, the test-by-examinee interaction would contribute about 30% to 50% of the variance and would likely be the largest error variance component. However, it is likely that this variance component would be higher in a sample of children with intellectual giftedness than in more normative samples due to the tendency to “regress toward the mean.” For example, children who obtain high IQs on one test are likely to obtain scores closer to the mean on subsequent occasions. Therefore, regression toward the mean is likely to influence the test-by-examinee variance component (Lohman & Korb, 2006; Bergeron & Floyd, 2013). Furthermore, based on previous research (Floyd et al, 2008; Irby & Floyd, 2011; Irby et al., 2013), it is hypothesized that the test-by-examinee interaction component for both samples would contribute the largest overall variance in IQs. Consistent with previous research (Irby & Floyd, 2011; Irby et al., 2013), it is also hypothesized that the other variance components would contribute negligible variance (i.e., <5%). Finally, the dependability coefficient is expected to fall below .80, which is the typical lower level boundary for acceptable reliability.

3. Third, how many children are meeting criteria (i.e., 120, 125, and 130 cut-scores) for intellectual giftedness on each of the three brief tests, and all possible combinations of tests?

Method

Participants

Participants included 36 children drawn from the population of third- through fifth-grade students attending a university campus school in an urban school district and receiving special education services for intellectual giftedness. The specific selection criteria for the children with intellectual giftedness included a previous psychoeducational assessment conducted within the past 5 years indicating they met the state special education eligibility criteria for intellectual giftedness. The sample included 36 children (23 boys, 13 girls) between the ages of 8 to 11 years (*M* = 9.5, *SD* = 0.9 years), and 28% were in 3rd grade, 22% were in 4th grade, and 50% were in 5th grade. In terms of race, 72% were White, 11% were Black, and 16% were otherwise classified (i.e., Asian/Pacific Islander or multiracial). None were of Hispanic origin and all spoke English as their primary language. Children's mean IQ from previous assessments was 128.4 (*SD* = 9.3, range = 111 to 151). Most children were administered the Reynolds Intellectual Assessment Scales (86%; RIAS; Reynolds & Kamphaus, 2003), 14% were administered other tests (i.e., WISC-IV; the Stanford Binet – Fifth Edition, SB-V, Roid, 2003; or the Wechsler Preschool and Primary Scale of Intelligence – Third Edition; WPPSI-III; Wechsler, 2002). In regards to cut-scores, 44% of participants met the 130 IQ criteria, 64% of participants met the 125 criteria, and 78% of participants met the 120 criteria. Only 8 children had IQs below 120. Children's mean achievement scores in math and reading from previous statewide assessments (with percentile rank values converted to standard scores with *M* = 100 and *SD* = 15) were 119.4 (*SD* = 9.8, range = 107 to 135) and 115.8 (*SD* = 7.4, range 104 to 129), respectively.

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 aged 4 to 90 years. The test takes about 15 to 30 minutes to administer and consists of three subtests. The KBIT-2 produces a Composite Intelligence Index (CIX) score from three subtests, Verbal Knowledge, Matrices, and Riddles. The CIX has a mean of 100 and a standard deviation of 15.The KBIT-2 is the revised version of the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990).

The KBIT-2 CIX score has demonstrated very strong internal consistency reliability across ages 6 to 11 (mean coefficient = .92). The CIX also has a very strong test–retest reliability (with an interval of 6 to 56 days between administrations) for the 4- to 12-year-old age range (*r* corrected for restriction of range = .88; Kaufman & Kaufman, 2004). Inter-rater reliability estimates have not been reported.

The KBIT-2 CIX has demonstrated satisfactory convergent validity based on strong to very strong correlations with IQs from full-length intelligence tests. A. Kaufman and N. Kaufman (2004) reported moderate to strong correlations between the KBIT-2 CIX and the Full Scale IQs (FSIQs) from the following full-length tests: the Wechsler Intelligence Scale for Children, Third Edition (WISC-III; Wechsler, 1991) and the WISC-IV. Specifically, the KBIT-2 CIX had moderate to strong correlations with the WISC-III FSIQ and WISC-IV FSIQ (*r* = .78 and .66, respectively). In the same samples, the KBIT-2 CIX sample mean was 3 ½ to 5 points lower than the WISC-III FSIQ sample mean and 2 points lower than the WISC-IV FSIQ sample mean.

A. Kaufman and N. Kaufman (2004) also reported two separate studies comparing the KBIT-2 CIX to IQs from brief intelligence tests. For example, the KBIT-2 CIX correlated strongly with K-BIT CIX for children ages 8 to 14 and moderately with the WASI FSIQ-4 for children ages 7 to 19 (*r* = .87 and .71, respectively). In the same samples, the KBIT-2 CIX sample mean was 2 points lower than the K-BIT CIX sample mean and the mean score difference was not shown for the WASI FSIQ-4.

Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II). The WASI-II (Wechsler, 2011) is an individually administered brief intelligence test designed for individuals ages 6 to 90 years. The WASI-II Full-Scale IQ (FSIQ-4) consists of four subtests, Vocabulary, Block Design, Similarities, and Matrix Reasoning. The mean for the FSIQ-4 is 100 and a standard deviation of 15. The WASI-II is the revised edition of the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999).

The WASI-II FSIQ-4 has very strong internal consistency in the overall child sample ages 6 to 16 (mean reliability coefficient = .96; Wechsler, 2011). The test–retest reliability of the FSIQ-4 (with an interval of 2 to 12 weeks between administrations) for ages 6 to 11 is very strong as well (*r* = .93). There appears to be little or no evidence of the inter-rater reliability of WASI subtest scores or its IQs; however, inter-scorer agreement is reported for all four subtests on 60 protocols. Overall, the WASI-II demonstrated very strong inter-scorer agreement (ranging from .94 to .99).

The FSIQ-4 has demonstrated satisfactory convergent validity based on correlations with IQs from full-length intelligence tests. There are two studies in the WASI-II Technical Manual (Wechsler, 2011) that compared scores from the WASI-II to those from full-length intelligence tests and a brief intelligence test. For example, the

WASI-II FSIQ-4 had strong to very strong correlations with the WISC-IV FSIQ and KBIT-2 CIX (*r* = .88 and .91, respectively). In the same samples, the WASI-II FSIQ-4 sample mean was similar to the WISC-IV FSIQ sample mean and was 2 points lower than the KBIT-2 CIX sample mean.

Woodcock–Johnson III (WJ III). The WJ III Tests of Cognitive Abilities (COG; McGrew & Woodcock, 2001) is an individually administered full-length intelligence test designed for individuals ages 2 to 90+ years. The WJ III COG has a Brief Intellectual Ability (WJ III BIA) measure formed from three subtests that take about 15 to 30 minutes to administer. The WJ III BIA score is derived from three subtests, Verbal Comprehension, Concept Formation, and Visual Matching. The WJ III BIA score has a mean of 100 and a standard deviation of 15.

The WJ III BIA has demonstrated very strong internal consistency reliability for ages 6 to 14 (mean reliability coefficient across ages 6 to 14 = .95). Test–retest reliability (with an interval of one day between administrations) was reported for only the Visual Matching subtest for ages 7 to 11 and (*r* = .87). There appears to be little or no evidence of the inter-rater reliability of WJ III BIA subtest scores or the WJ III BIA itself.

The WJ III BIA has demonstrated satisfactory convergent validity based on moderate to strong correlations with IQs from full-length intelligence tests. McGrew and Woodcock (2001) described studies in which the WJ III BIA (obtained after full administration of the WJ III COG) was compared to scores from six full-length intelligence tests. For example, the WJ III BIA was moderately correlated with the WISC-III FSIQ and Differential Ability Scales, General Conceptual Ability (DAS, GCA; Elliott, 1990) and the Cognitive Abilities Scale Full-Scale Score (CAS FSS; Naglieri &

Das, 1997; *r* = .69, .70, and .70, respectively). In the same samples, the WJ III BIA sample mean was 4 points lower than the WISC-III FSIQ sample mean, 3 points lower than the DAS GCA sample mean, and 5 points lower than the CAS FSS sample mean. The WJ III BIA was also compared to the WJ III COG General Intellectual Abilities (GIA), which stems from performance on the three subtests contributing to the WJ III BIA and four other subtests. Not surprisingly, their correlation was very strong (*r* = .92).

Procedures

Recruitment and selection of participants. Third, fourth, and fifth graders who were classified as intellectually gifted and were currently receiving special education services through the Creative Learning in a Unique Environment (CLUE) program were invited to participate. The "Basis for Selection" section of the Informed Consent Form (see Appendix B) included the statement; "All third through fifth grade children in your child's school who receive special education services through the CLUE program have been invited to participate in this study." CLUE teachers distributed recruitment letters (see Appendix A) intended for the parents of these children. These letters included an explanation of the study (see Appendix A), an Informed Consent Form (see Appendix B) for parents to keep for their records, an Informed Consent Form for parents to sign and return, a Signature Form (see Appendix C, Part 1) that indicated parents read and understood the Informed Consent Form allowing their children to participate in the study, and a Demographics Form (see Appendix C, Part 2). Parents were asked to return the signed Informed Consent Form, the Signature Form, and the Demographics Form to the teachers in a sealed envelope. Teachers provided the unopened envelopes from parents to the researchers. Overall, forms were sent home to parents of 58 children, and 36 parents (62%) returned forms allowing their children to participate.

After parental consent is obtained, the researcher provided a list of participants in the study to the CLUE teachers, and CLUE teachers provided the researchers with the results from each student’s most recent cognitive and achievement assessment (see Appendix D). This information was used to ensure that the students have not completed any of the tests used in this study within the last 12 months and to enhance the description of the sample. Access to this information was communicated to parents in both the Letter of Invitation (see Appendix A; which includes the statement; “The results from your child’s most recent cognitive and achievement tests will be shared with the examiners to make sure that they do not take the same tests again”) and the Informed Consent Form (which includes the statement; “Demographic information and previous assessment results also will be obtained from your child’s CLUE teacher”). None of the participants were excluded or deemed ineligible for this study due to similar testing within the past 12 months.

Post-consent contact and scheduling of sessions. Prior to each initial testing session, the parents of the child were contacted by phone by the author in order to schedule a time for them to participate in the study during the hours during the school day in which children were typically participating in CLUE sessions or after school (e.g., 3:00 to 5:00). About 58% of testing sessions were conducted during CLUE sessions, and the rest were conducted after school. When scores from sessions during CLUE time and after school were contrasted, there were no statistically significant differences across IQs, *F*(1, 106) = 0.56, *p* = .46, or for any single IQ, KBIT-2 CIX, *F*(1, 34) = 0.10, *p* = .75; WASI-II

FSIQ-4, *F*(1, 34) = 2.245, *p* = .14; or WJ III BIA, *F*(1, 34) = 0.02, *p* = .89. Thus, no confounds related to the time of testing on the participants' IQs were evident.

After scheduling of the testing session, the participants were randomly assigned to an examiner in order to control for individual differences in examinees across the "primary examiners." To accomplish this goal, each examiner was assigned a number between 1 and 7 using a random list generator, and using a random number list, examiners were matched to participants. There were at least three examiners available during each testing session timeslot, which was done in case the initially selected examiner was unable to attend the scheduled session. In cases of random assignment of an examiner to a participant with a scheduled assessment session that the examiner could not attend, another examiner was selected using the same methods. This process was completed until an examiner was scheduled for each appointment timeslot.

Testing sessions. If a child's parent agreed to the stipulations included in the Consent Form (see Appendix B), the child was introduced to the study by a primary examiner. Once the child understood the study, they provided assent to participate (see Appendix E). Each participant who assented completed all tests with the primary examiner in one test session, except for six participants who were administered tests across two testing sessions due to time constraints (e.g., testing longer than 2 hours or the child needed to go to lunch). In these six cases, a follow-up testing session was scheduled within the next 7 days. In order to control for carryover effects (e.g., practice effects and fatigue), all tests were administered in a counterbalanced order. After the testing session, each child chose an incentive (e.g., a pencil, an eraser, or a "silly-band"). Additionally, a

$10 store gift card was distributed to the child by envelope within one week of the testing session; it was sent home with the child with the parent’s name on the envelope.

Each test session was video recorded, and responses were independently scored by the “secondary examiner.” To ensure that subtests involving visual stimuli or manipulatives (e.g., blocks) were scored accurately by the author, a video camera was aimed at the testing table to record these responses (see Appendix F). Both primary and secondary examiners obtained norm-referenced standardized scores using scoring software (WJ III) or norms tables included in test manuals (KBIT-2 and WASI-II).

Primary examiners. The brief intelligence tests were administered to participants by one of seven examiners (i.e., “primary examiners”). The primary examiners each passed two graduate-level assessment courses and a graduate-level assessment practicum and completed two training sessions with the author prior to administering tests. Each primary examiner completed a demographics survey (see Appendix G). The primary examiners reported completing 28 to 130 hours of graduate coursework (*M* = 82.4 hours) and 300 to 2,300 graduate practicum hours (*M* = 1285.7 hours) prior to administering the tests. They reported administering 11 to 600 comprehensive and screening tests (*M* = 145 tests). All primary examiners were White; with 4 women and 3 men.

Training sessions for primary examiners consisted of reviewing administration and scoring procedures for the three brief intelligence tests included in the study as well as direct instruction in how to use electronic recording equipment. After the first training session, each primary examiner submitted protocols for each test in order to ensure competence (i.e., fewer than two invalidating errors across all three tests) in administration and scoring. The protocols were reviewed by the author to ensure that no invalidating errors were present, and minor errors were discussed with the primary examiners during the second training session. Each primary examiner demonstrated competency on each test prior to data collection. Six primary examiners completed an equal number of assessments (i.e., five assessments each); one primary examiner completed six assessments.

The primary examiners completed the scoring within one week of the test administration. Primary examiners were allowed 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 not able to consult with the author (or faculty or professional supervisors involved in the study), who remained blind to the results from the tests. The author reviewed the session recordings months later. After each administration, primary examiners placed completed protocols in folders in a filing cabinet that were monitored by the faculty advisor to ensure the secondary examiner remained blind before scoring.

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., the Vocabulary subtest of the WASI-II) and when unsure of how to score responses. Scoring issues were discussed with the faculty advisor. For cases in which responses were inaudible or the primary examiner demonstrated an administration error (e.g., administering an item incorrectly, scoring an item incorrectly, and failing to establish a floor or ceiling), a list of random numbers were consulted to determine whether or not to award credit for items that were affected. For example, if the primary examiner administered an item incorrectly, the secondary examiner reviewed the random numbers list in order to decide whether to award credit.

Results

Data Screening and Tests of Assumptions

Preliminary data analyses were conducted with each of the three tests for each examiner to ensure that the assumptions of multivariate analysis and correlations were not violated (Tabachnick & Fidell, 2013). There were missing values for two participants for the secondary examiner due to recording equipment failure (i.e., the camera battery died midway through an assessment and a camera memory card was full). There were no univariate (zs < │3.29│) or multivariate outliers found for any of these variables. No IQ was notably skewed for either examiner (all values < │1.0│) or had notable kurtosis for either examiner. All values were less than │1.0│ except for the KBIT-2 CIX and WJ III BIA for the primary examiner (kurtosis = 1.50 and 1.17, respectively). All other assumptions of paired-samples *t*-tests were judged not to be violated

Table 1 includes the means and standard deviations for each IQ, subtest, and composite by examiner. The means ranged from 115.31 (WJ III BIA) to 123.80 (KBIT-2 CIX) for the primary examiner and from 116.00 (WJ III BIA) to 123.82 (KBIT-2 CIX) for the secondary examiner. The means for both examiners were at least one standard deviation above 100. The standard deviations ranged from 10.00 (KBIT-2 CIX) to 13.17 (WJ III BIA) for the primary examiner and from 9.95 (KBIT-2 CIX) to 12.09 (WJ III BIA) for the secondary examiner. The standard deviations for both types of examiners were less than 15 in every case, which indicates restriction of range of the samples, which was expected due to their being previously identified with intellectual giftedness).

Table 1

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

	Primary examiner			Secondary examiner			Inter-rater reliability			
IQs	*M*	*SD*	Range	*M*	*SD*	Range	r*	r_c	*M* diff	*t*
KBIT-2 CIX	123.8	10.0	95-149	123.8	9.9	105-149	.96	.98	-0.02	-0.87
WASI-II FSIQ-4	118.2	11.6	92-144	117.2	10.2	100-142	.93	.96	0.98	1.95
WJ III BIA	115.3	13.1	75-141	116.0	12.0	85-139	.96	.97	-0.69	1.63
Subtests/ Composites	*M*	*SD*	Range	*M*	*SD*	Range	*r*	r_c	*M* diff	*t*
KBIT-2 Verbal	122.9	10.4	91-148	123.7	10.1	97-147	.95	.97	-0.77	-1.38
KBIT-2 Non-verbal	118.1	11.0	92-141	117.2	11.3	92-141	.95	.97	0.90	-0.49
WASI-II Block Design[a]	56.3	9.8	36-78	56.8	9.5	33-79	.96	.96	-0.55	-0.35
WASI-II Vocabulary[a]	66.9	7.0	54-80	65.2	5.6	55-80	.59	.72	1.71	1.94
WASI-II Matrix Reasoning[a]	60.6	10.2	40-80	60.3	9.5	40-80	.99	.99	0.32	2.61
WASI-II Similarities[a]	55.3	10.7	38-80	54.3	8.0	41-80	.83	.81	1.01	-0.83
WASI-II VCI	117.8	12.2	99-154	115.5	8.1	99-138	.81	.86	2.33	1.96
WASI-II PRI	114.8	15.7	87-152	115.2	15.9	87-154	.98	.98	-0.43	-0.47
WJ III Test 1	113.0	9.9	75-132	115.2	9.3	82-136	.88	.94	-2.17	-2.57
WJ III Test 5	115.2	13.4	88-142	115.3	13.8	92-141	.92	.93	-0.15	0.39
WJ III Test 6	103.8	15.8	70-133	104.2	16.0	72-140	.99	.99	-0.42	-0.97

*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-II = Wechsler Abbreviated Scales of Intelligence, Second Edition; WJ III COG = Woodcock–Johnson III Tests of Cognitive Abilities; WJ III BIA = Brief Intellectual Ability; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index.

[a]Scores are age-based *T*-scores (*M* = 50, *SD* = 10).

*All correlations significant at $p < .001$ (two-tailed).

Convergent Validity and Mean Differences across Tests for Brief IQs

To examine the convergent validity evidence supporting the IQs, correlations for each test with the other two tests for both examiners were conducted, resulting in six 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 2 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 (Floyd et al., 2008). Results presented above the diagonal in Table 2 reveal one moderate correlation (between the KBIT-2 CIX and the WASI-II FSIQ-4) and two strong correlations for the primary examiners. For the secondary examiner (see below the diagonal), one correlation was strong (between the KBIT-2 CIX and the WJ III BIA), and the remaining two correlations were moderate. The bottom of Table 2 includes correlations between the IQs corrected for range restriction. For both the primary and secondary examiners, there was one very strong correlation (between the WJ III BIA and the KBIT-2) and two strong correlations, which supports the first hypothesis that the relations across tests would be moderate to strong.

Table 2
Correlation Matrix for Tests by Examiner

	Obtained correlations		
Measure	1	2	3
1. KBIT-2 CIX	-	.47	.75
2. WASI-II FSIQ-4	.62	-	.60
3. WJ III BIA	.77	.70	-
	Corrected correlations		
Measure	1	2	3
1. KBIT-2 CIX	-	.81	.90
2. WASI-II FSIQ-4	.85	-	.84
3. WJ III BIA	.91	.83	-

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. KBIT-2 = Kaufman Brief Intelligence Test, Second Edition; WASI-II FSIQ-4 = Wechsler Abbreviated Scale of Intelligence, Second Edition Full-Scale IQ-4 WJ III BIA = Woodcock–Johnson III Tests of Cognitive Abilities Brief Intellectual Ability.
*All correlations significant at $p < .001$ (two-tailed).

A one-way ANOVA was used to evaluate possible order effects on the IQs from the intelligence tests administered first, second, and third—regardless of which test it was. Results were nonsignificant ($p > .10$); thus, counterbalancing eliminated any order effects that may have been present. Another 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, $F(2, 105) = 4.99$, $p = .01$. For the secondary examiner, there was also a significant effect of test on IQs, $F(2, 100) = 5.19$, $p = .01$. Tukey post-hoc results indicated that there were significant differences between the KBIT-2 CIX and the WJ III BIA for the primary examiner and between the KBIT-2 CIX and both the WASI-II FSIQ-4 and the WJ III BIA ($p < .05$) for the secondary examiner. All other comparisons were nonsignificant ($p > .10$).

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

To examine the inter-rater reliability of the IQs, one correlation was calculated between the IQs from both examiners from each test, resulting in a total of three correlations. 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. Paired-samples *t*-tests were also conducted to evaluate the mean differences in IQs between examiners scoring each test. On the left side of Table 1, uncorrected correlations and corrected correlations across examiners, mean differences across examiners, and the results of paired-samples *t* tests for each IQ and (for reference) each subtest and composite score are presented. Uncorrected inter-rater reliability coefficients ranged from .96 (for both the KBIT-2 and the WJ III BIA) to .93 (for the WASI-II FSIQ-4). After correcting for range restriction in scores, the inter-rater reliability corrected coefficients ranged from .98 (KBIT-2) to .93 (WASI-II). Based on Nunnally and Bernstein (1994), inter-rater reliability for IQ tests should be considered adequate when $r > .90$. The uncorrected and corrected correlations for each test met this standard, and the corrected correlations are strong to very strong, which supported the first hypothesis that relations across examiners would be moderate to strong. The scoring items on the WJ COG III and the KBIT-2 require less examiner judgment when scoring than the WASI-II, and they produced the strongest correlations in IQs across examiners. Mean differences across IQs and examiners were approximately 1 standard score point more or less, and all *t*-tests revealed nonsignificant mean differences between IQs, *p*s > .05.

Dependability Analysis

Finally, IQs were entered into a G-theory analysis to examine their dependability. Variance components were computed using PASW 18.0, and dependability coefficients (a.k.a., phi coefficients) were calculated to provide overall dependability (Brennan, 2001; Shavelson & Webb, 1991). The variance estimate attributable to differences across IQs was considered universal score variance and used as the numerator in the formula to calculate the dependability coefficients. The variance estimates attributable to test, examiners, all interactions, and 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 universal score and error variance.

Table 3 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 approximately two-thirds of the variance; in fact, it accounted for 57% of the variance. Thus, the remainder of variance was due to systematic or random error.

Table 3

Variance Component Estimates and Absolute Dependability Coefficients

	Estimated variance components	
Facet	Brief or abbreviated IQs	Percent of variance
Examinee	81.392	57%
Examiner [a]	0	0%
Test	17.105	12%
Examinee-by-examiner	1.878	1%
Examinee-by-test	38.309	27%
Examiner-by-test	0.449	0%
Residual	4.59	3%
Total	143.723	
ϕ	.80	

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

When considering error variance components, the largest proportion of variance was attributed to the test-by-examinee interaction; it accounted for 27% of the variance. These results do support the hypothesis that the test-by-examinee interaction component would contribute the largest error variance in IQs. The test characteristics contributed 12% of the variance, supporting the hypothesis that the test component would contribute small but notable variance in IQs and exceed that in previous studies, which is likely due to the KBIT-2 producing significantly higher mean scores than the WJ III for both examiners and the WASI-2 for the secondary examiner (see Table 1).

The examiner, examinee-by-examiner interaction, and examiner-by-test interaction contributed minimal variance in scores. These results supported the hypotheses that these components would be minimal. Residual variance was only 3%. The dependability coefficient was .80, indicating suspect dependability of IQs across brief intelligence tests and examiners, the two most well-studied and powerful influences on score differences.

Partial models. Several additional analyses were conducted to better understand the source of sizable error variance components. First, three partial models were analyzed, with IQs from one intelligence test omitted from each. In the model omitting the KBIT-2 CIX, the test variance was reduced from 12% to 0%, whereas the size of this variance component either increased or was virtually unchanged in the other partial models. Thus, the test variance in the full model was due—in very large part—to the inclusion of the KBIT-2 CIX in the analysis. In the partial analysis omitting the WASI-II FSIQ-4, the test-by-examinee variance was reduced from 27% to 19%, whereas the size of this variance component changed little in the other partial models. Thus, the WASI-II

FSIQ-4 contributed substantial variance to the interaction between the individual and the test.

Cut-off Analysis

To examine if the participants in this study would meet criteria for intellectual giftedness (e.g., 130, 125, or 120 cut-scores) based on scores from the primary examiner (seeing that there was no variance in IQs attributable to the scorer and related facets), a series of steps were followed. First, the 90% and 95% confidence intervals for each test and age range were calculated (Kranzler & Floyd, 2013). First, the standard error of measurement was calculated for each IQ by subtracting the median internal consistency reliabilities for the age-groups included in the sample (as reported in the Method section), obtaining the square root of this difference, multiplying this value by 15 (the standard deviation of the IQs). Finally, the standard error of measurement for each IQ was multiplied by constants (1.65 for the 90% confidence interval and 1.96 for the 95% confidence interval) to obtain the values for the confidence intervals. The 90% and 95% confidence intervals for the KBIT-2 were +/- 7.13 and +/- 8.47, respectively. The 90% and 95% confidence intervals for the WASI-II were +/- 4.95 and +/- 5.88, respectively. The 90% and 95% confidence intervals for the WJ III BIA were +/- 5.53 and +/- 6.57, respectively. Half of each confidence interval value (e.g., 7.13 and 8.47 for the KBIT-2 Composite) was added to the obtained IQs for each test for the primary examiner to create new variables. Additional dichotomous variables were then created for each cut-score (i.e., 130, 125, and 120) to determine the number of participants who met the cut-score criteria for each test and every combination of tests. For example, for the KBIT-2 90% confidence interval, a new variable was created by adding 7.13 to each IQ. Three additional variables were created by converting scores above each cut-score (i.e., 120, 125, and 130) into dichotomous variables. In order to determine if a child met cut-score criteria on two or more tests, their dichotomous variables were added together and a frequency distribution used to determine if they met criteria on both tests. For example, if after adding the two dichotomous variables together the child had a score of 2, then they met the cut-score criteria on both tests.

Table 4 shows the numbers and percentages of participants who met the various cut-score criteria for each test and combination of tests at both the 90% and 95% confidence intervals. The KBIT-2 CIX was almost twice as likely to result in scores above the 130 cut-off as both the WJ III BIA and WASI-II FSIQ-4. Overall, it appears that the KBIT-2 CIX results in more children meeting or exceeding each of the cut-offs, and almost all of them (92%) obtaining an IQ above 120 for both the 90% and 95% confidence intervals. Using a 90% confidence interval, only 9 participants (25%) met the 130 criteria on all three tests, 10 participants (28%) met the 125 criteria on all three tests, and 16 participants (44%) met the 120 criteria on all three tests. Similar results were observed at the 95% confidence interval. Only 9 participants (25%) met the 130 criteria on all three tests, 11 participants (31%) met the 125 criteria on all three tests, and 18 participants (50%) met the 120 criteria on all three tests.

Table 4

Numbers and Percentage of Children Exceeding Cut-off Scores Using Primary Examiner Data

	Cut-Offs for 90% Confidence Intervals					
Measures	130		125		120	
	#	%	#	%	#	%
KBIT-2 CIX	21	58%	31	86%	33	92%
WASI-II FSIQ-4	12	33%	15	42%	22	61%
WJ III BIA	11	31%	14	39%	18	50%
	Cut-Offs for 90% Confidence Intervals					
Pairwise combinations	130		125		120	
	#	%	#	%	#	%
KBIT-2 & WASI-II FSIQ-4	10	28%	13	36%	21	58%
KBIT-2 & WJ III BIA	11	31%	14	39%	18	50%
WASI-II FSIQ-4& WJ III BIA	9	25%	10	28%	16	44%
All Tests	9	25%	10	28%	16	44%
	Cut-Offs for 95% Confidence Intervals					
Measures	130		125		120	
	#	%	#	%	#	%
KBIT-2 CIX	23	64%	31	86%	33	92%
WASI-II FSIQ-4	12	33%	17	47%	24	67%
WJ III BIA	11	31%	16	44%	21	58%
	Cut-Offs for 95% Confidence Intervals					
Pairwise combinations	130		125		120	
	#	%	#	%	#	%
KBIT-2 & WASI-II FSIQ-4	10	28%	15	42%	23	64%
KBIT-2 & WJ III BIA	11	31%	16	44%	21	58%
WASI-II FSIQ-4& WJ III BIA	9	25%	11	31%	18	50%
All Tests	9	25%	11	31%	18	50%

Note. All results are based on IQs obtained from the primary examiners. KBIT-2 = Kaufman Brief Intelligence Test, Second Edition; WASI-II FSIQ-4 = Wechsler Abbreviated Scale of Intelligence, Second Edition Full-Scale IQ-4; WJ III BIA = Woodcock–Johnson III Tests of Cognitive Abilities Brief Intellectual Ability.

Discussion

High-stakes assessments are required to determine if a child is intellectually gifted; therefore, it is important to be mindful of several important issues related to accepting an IQ from a single test as valid. Moreover, it is important to understand the exchangeability of IQs for children with intellectual giftedness. However, recent IQ exchangeability studies have primarily focused on nonclinical samples (Floyd et al., 2008; Irby & Floyd, 2011; Irby et al., 2013). This study furthers the study of IQ exchangeability by addressing three weaknesses in prior research. First, it evaluated the overall dependability of brief IQs for children with intellectual giftedness. Second, it helped determine the exact reasons for suspect IQ exchangeability in a gifted sample. Third, it also determined what percentage of children who meet criteria for intellectual giftedness would likely meet it on subsequent tests.

Frequently, correlations between IQs and mean differences across IQs have been used by researchers to examine the effects of test characteristics in isolation and the effects of examiners in isolation. This study was different from most prior studies because it examined the effects of both the inter-rater reliability, which has only minimally been studied (e.g., Irby & Floyd, 2011; Irby et al., 2013), and convergent validity across intelligence tests collectively. It was important to further evaluate these influences because it helped determine their total effect on exchangeability of IQs as well the strength of their effects. Furthermore, due to time-constraints in administering full-length intelligence tests and associated confounds that may weaken score exchangeability, the study employed brief intelligence tests in order to complete the testing in one sitting.

Dependability and Exchangeability of Brief IQs

When test characteristics, examiner effects, 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., .92 to .96), the dependability coefficient was .80, which did not meet the recommendation of .90 when important decisions are being made (i.e., assessment of intellectual giftedness; Nunnally & Bernstein, 1994). The dependability coefficient in this study is higher, however, than what Irby and colleagues (2011, 2013) found when examining IQs from four brief intelligence tests in a sample of college students (dependability coefficient = .73).

Additionally, the dependability coefficient found in this study is higher than the mean value reported by Floyd et al. (2008) when examining IQs from pairs of full-length intelligence tests (*M* dependability coefficient = .73). The dependability value was expected to be far lower in this study due to the use of brief intelligence tests, which are less reliable than full-length tests. In addition, this study examined an additional facet of error variance (i.e., examiner effects) and compared three tests versus two tests, which could have produced a lower dependability coefficient than those found in Floyd et al. However, in Floyd et al., more than half (6) of the 10 pairwise IQ comparisons yielded dependability coefficients in the same general range (.51 to .93; *M* = .73) as this study. It is probable that administering all tests in counterbalanced order in a single session in the current study and in the Irby and colleagues (2011, 2013) study contributed to slightly higher dependability coefficients than that from the multiple IQ comparison analysis reported by Floyd et al.

Test characteristics. The influence of test characteristics on the exchangeability of IQs was assessed through correlations between IQs from different tests, a one-way ANOVA comparing IQs for each test, and a G-theory analysis. In regards to convergent validity indicated by the correlations between IQs from different tests, the hypothesis that there would be moderate to strong correlations was supported. These results are also similar to the results from Irby and Floyd (2011). Based on the results from the one-way ANOVA, there was a significant effect of the tests for both examiners, which indicates potential problems with exchangeability. More specifically, the KBIT-2 CIX was significantly higher than at least one of the other tests for the primary and secondary examiners. These results do not appear to support the Flynn effect—individuals would receive IQs that are higher on tests normed earlier (Flynn, 2006, 2009)—because the KBIT-2 normative data were collected from 2001-2003 (Kaufman & Kaufman, 2004), whereas WJ III normative data were collected from 1996-1999. Therefore, in order to support the Flynn effect, the WJ III should have produced significantly higher IQs than the KBIT-2 and WASI-II (normative data was collected from 2010-2011).

In a manner similar to the results from previous exchangeability studies (e.g., Floyd et al., 2008; Irby & Floyd, 2011; Irby et al., 2013), the G-theory analysis revealed that the test component contributed 12% of the variance in brief IQs in this study. This value was slightly higher than previous studies and supported the second hypothesis (i.e., that test variance would contribute approximately 5-10% of the variance). This relatively small percentage of variance (and the ANOVA results) may be attributable to the testing format employed by various subtests composing the different IQs. Furthermore, when the KBIT-2 CIX is removed from the analysis, the variance in IQs due to the test is reduced to 0%. In addition, more children were able to meet cut-score criteria on the KBIT-2 CIX than the other two tests. For example, children performed approximately ½ of a standard deviation better on the Verbal composite on the KBIT-2 than on WASI-II verbal subtests (i.e., Vocabulary and Similarities) and the WAIS-II VCI as well as the WJ III Verbal Comprehension subtest (see Table 1). Furthermore, children were more likely to obtain

IQs at or above 120 on the KBIT-2 (92%) than the WASI-II (61%) or WJ III (50%), which was likely due to the higher Verbal composite scores on the KBIT-2.

Examiner effects. Results of the G-theory analysis showed that the variance due to the examiner and the component representing interactions with examiner effects contributed negligible variance in IQs. Furthermore, the results from the inter-rater reliability analysis were congruent with the G-theory analysis. For example, the corrected inter-rater reliability for each test ranged from .96 to .98 (M = .97), which is well above the .90 criterion offered by Nunnally and Bernstein (1994). In support of the second hypothesis, the G-theory and inter-rater reliability analyses indicate that examiner's scoring differences are not major confounds (on a relative scale) to intelligence test score interpretation. Therefore, despite the potential for examiner error, it does not appear to have a significant effect on the variation in IQs.

Test-by-examinee effects. Irby and colleagues (2011, 2013) found the largest error variance component to be due to the test-by-examinee interaction. Similarly, in support of the hypothesis regarding the largest error variance component, the test-by-examinee interaction in this study accounted for 27% of the variation in IQs. In short, some students performed well on some tests, whereas others performed better on other tests.

The content of items, cognitive processes evoked, and the required response modalities across subtests contributing to IQs may vary substantially across intelligence tests. It is possible that the variation in content, processes, and responses, as well as the examinees' varying reactions to them appears to cause different IQs across tests (McGrew, 2009). 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, which is less likely to affect variation in IQs for children of average intellectual functioning.

In addition, it is possible that, as hypothesized, the higher variance attributed to test-by-examinee interaction is related to "regression toward the mean," which means that children are likely to obtain scores closer to the mean on subsequent occasions and, therefore, would be difficult for all three IQs to exceed a specific cut-score across multiple intelligence tests (as evidenced in the cut-score analysis). Moreover, the cut-score analysis showed that only 25% of participants were able to obtain scores above 130 across all three tests at both the 90% and 95% confidence intervals.

Limitations

Due to the nature of the sample, results can be generalized to the population of children with intellectual giftedness; thus, they are helpful in understanding the influences on IQs for eligibility testing for children with intellectual giftedness. However, intelligence tests are administered more frequently to children in order to determine if they meet criteria for an intellectual disability, which is at the other extreme of the normal curve. In addition, children usually receive a full-length intelligence test as part of a comprehensive psychoeducational assessment instead of a brief intelligence test. Therefore, the results of this study may not generalize to other populations that require testing.

This study employed an innovative method to examine examiner effects without the added confound of test-retest, as well as controlling for differences in examiner-examinee rapport. Although this innovative method made it possible to examine test characteristics and examiner effects in a single study, the procedure of recording and later scoring responses to test items was not perfect and contributed difficulty to scoring some items. 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 for Practice and Future Research

The results from this study suggest that there is suspect dependability in IQs from brief intelligence tests, especially when used for making high-stakes decisions (e.g., special education eligibility). This conclusion is also supported by the relatively low agreement in meeting the cut-scores on each test and combination of tests. Based on these results, examiners should be cautious in using brief intelligence tests to identify children for special education because a child may or may not qualify for services depending on which test is used. Instead, brief intelligence tests should be primarily used for screening purposes. Additionally, lower cut-scores should be used when deviant scores (e.g., scores more than one standard deviation above the mean) are used as part of the identification process.

Results suggest that the KBIT-2 CIX produces somewhat inflated scores—at least compared to the two other tests—and should be used with caution. Consistently, across multiple studies (e.g., Floyd et al., 2008; Irby & Floyd, 2011; Irby et al., 2013), the test-by-examinee interaction (or pilot-by-altimeter or recording method-by-runner interaction) contributed the largest amount of error variance in IQs. For this reason, examiners must be careful in choosing a test that best assesses an individual's abilities (e.g., choosing a test without processing speed subtests for individuals who are slow responders) and should use full-length tests when making high-stakes decisions.

The variance due to examiner and all interactions with the examiner resulted in negligible differences between IQs. It is possible that these negligible differences may be due to the more explicit scoring procedures in recent tests. Despite this minimal variance, practitioners and trainers should remain steadfast in striving to reduce subjectivity in scoring. In addition, 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. For example, during the standardization process for intelligence tests, a small sample should be administered the test twice with a different examiner each time. However, due to confounds (e.g., maturation), a method similar to the one utilized in this study (i.e., recording sessions and having a secondary examiner score blank protocols and adjust scores for examiner error) may be more appropriate. Moreover, the majority of current research focuses on inter-scorer agreement instead of true 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 or intellectual and developmental disabilities.

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).

Bergeron, R., & Floyd, R. G. (2013). Individual part score profiles of children with intellectual disability: A descriptive analysis across three intelligence tests. *School Psychology Review, 42,* 22-38.

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.

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

Elliott, C. (2007). *Differential Ability Scales, Second Edition*. San Antonio, TX: Psychological Corporation.

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

Irby, S. M., & Floyd, R. G. (2011). *Exchangeability of Brief Intelligence Tests: Illuminating the Influence on Error Variance Components on IQs.* Master's thesis.

Irby, S. M., Floyd, R. G., & Bergeron, R. (2013). *An Analog Study of the Exchangeability of Brief Intelligence Tests: Illuminating the Influence of Error Variance Components on IQs.* Unpublished manuscript.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Preager.

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

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

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

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted, 29,* 451-484.

Marland, S. (1972). Education of the gifted and talented: Report to Congress. Washington, DC: U. S. Government Printing Office.

McClain, M-. C., & Pfeiffer, S. (2012). Identification of gifted students in the United States today: A look at state definitions, policies, and practices. *Journal of Applied School Psychology, 28*, 59-88.

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

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

Naglieri, J. A., & Das, J. P. (1997). *Cognitive Ability Scale.* Itasca, IL: Riverside.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed., pp. 264-265). New York, NY: McGraw-Hill.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scale.* Odessa, FL: PAR Inc.

Robinson, N. M. (2005). In defense of a psychometric approach to the definition of academic giftedness. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 280-294). New York, NY: Cambridge University Press.

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.

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

Stephens, K. R. (2011). Federal and state response to the gifted and talented. *Journal of Applied School Psychology*, *27*(4), 306-318.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Allyn & Bacon.

Torrance, E. P. (1974). *Torrance Tests of Creative Thinking.* Bensenville, IL: Scholastic Testing Service.

United States Department of Education. (2000). *Office for Civil Rights (OCR) Elementary and Secondary School Survey.* Washington, DC: Author.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children, Third Edition.* San Antonio, TX: Pearson.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale, Third Edition.* San Antonio, TX: Pearson.

Wechsler, D. (1999). *Wechsler Abbreviated Scales of Intelligence.*San Antonio, TX: Pearson.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence, Third Edition.* San Antonio, TX: Pearson.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children, Fourth Edition*. San Antonio, TX: Pearson.

Wechsler, D. (2011). *Wechsler Abbreviated Scales of Intelligence, Second Edition manual.* San Antonio, TX: Pearson.

Worrell, F. (2009). Myth 4: A single test score or indicator tells us all we need to know about giftedness. *Gifted Child Quarterly, 53*, 242-244.

Appendix A
LETTER OF INVITATION

Dear Parent or Guardian:

Your child's principal has agreed to allow students from your child's school to participate in a study to better understand the thinking abilities of children with giftedness and are currently receiving special education services through the Creative Learning in a Unique Environment (CLUE) program. We are asking your help with this study.

If you agree to allow your child to participate in our study, your child will complete a series of tasks measuring thinking abilities on a one-to-one basis at school after school hours. As part of this study, the results from your child's most recent cognitive and achievement tests will be shared with the examiners to make sure that they do not take the same tests again. Your child's name and performance on the tasks will be kept confidential within the limits allowed by law, and participation is voluntary. Your child will not be placed in any harm by taking part in our study.

Because we recognize and appreciate the notable time and effort required of you, your child, and your child's school, we want to thank you all for participating in this project. We want to thank parents by providing a $10 giftcard to a local store for allowing their children to participate. Upon completion of your child's participation in the project, the giftcard will be sent home with the child.

If you are willing to allow your son or daughter to participate in our study, complete the following steps:

1. Carefully read the Informed Consent Form, which explains all of the details of the study. Keep one copy for your files and sign the second copy.

2. Complete the Signature Form to indicate that you give your child permission to participate in this project.

3. Complete the Child Information Sheet.

4. Return the signed Informed Consent Form, the Signature Form, and the Child Information Sheet in the enclosed envelope with your child to school.

If you would like more information about the study before allowing your son or daughter to participate, please contact me in the Psychology Department at The University of Memphis at (901) 340-7212. We hope you are willing to work with us. Thank you for your time and consideration in this matter.

Sincerely,

____

Sarah Irby, M.S. .
Primary Investigator

Appendix B
PARENT CONSENT FORM

Your child has been invited to participate in a research project entitled, **Exchangeability of Brief Intelligence Tests: Illuminating Error Variance Components' Influence on IQs for Children with Intellectual Giftedness**. The purpose of the study is to investigate a group of cognitive tests that measure a variety of thinking abilities. These tests have already been evaluated based on assessment of many thousands of children and adults in the United States. The research will be examining these test scores to see their similarities and differences. As part of this project, your child will complete a series of brief tasks assessing thinking and memory skills in a one-on-one setting with a trained examiner. ***We are asking your permission to include you and your child in our research project.*** Testing sessions will be video recorded and will last approximately 1 to 2 hours. We will strive to avoid including any identifying information (i.e., your child's name or face) on the video recording, which will be focused at the testing materials. The recordings will be destroyed after use (within 18 months). ***These assessments will be completed at your child's school after school hours (at an agreed upon time) in the fall (August-December) of this year.***Additionally, previous assessment results will be obtained from your child's CLUE teacher to make sure they do not take the same tests again.Please provide your telephone number below so that we may contact you to schedule a testing session that is convenient for you and your child. This study will help us understand the measurement of the thinking abilities of gifted students. You will receive a $10 giftcard to a local store for your child participating in this study.

Research materials and data related to your child will be kept confidential within the limits allowed by law. Any reporting of results will not identify the school or any students. The University of Memphis does not have a fund set aside for compensation in the case of study related injury, although there are no more than minimal risks associated with participation in the study. Participation in this study is completely voluntary and you are free to withdraw your child from the study at anytime without giving a reason and without penalty. You have the right to view the results of this study, regardless of whether your child completes all study-related measures. You can also have your child's information removed from the research record or destroyed. This study should benefit students by providing valuable information about how intellectual giftedness is measured.

Please feel free to ask any questions of the investigator before beginning the study and at any time during or after completion of the study. If you have any questions about the project, please call Sarah Irby (901) 340-7212. For additional information regarding your rights as a research participant, please contact Jacqueline Y. Reid, Administrator for the University of Memphis' Institutional Review Board for the Protection of Human Subjects, at 901-678-2533.

Thank you for your time and effort.

Principal Investigator:

____

Sarah Irby, M.S.
Department of Psychology
University of Memphis
Memphis, TN 38152
Email: sarahmirby@gmail.com

Appendix C
Part 1
SIGNATURE FORM

Check the appropriate box below indicating whether you wish your child to participate in the reading study. Fill out the information requested, place in the envelope provided, and return this completed form to your child's teacher.

I have read the information in the consent form and understand my rights and my child's rights as a research participant. I understand that I may contact the investigators to answer questions before allowing my child to participate. Refusal to participate will involve no penalty or loss of benefits to which I am otherwise entitled.

I □ My child and I do want to participate.

□ I would like more information, please call me.

Your name (print):

____
First name *Last name*

____
Telephone # (Will be used to schedule testing session)

________________________________ ________________
Signature of Parent or Legal Guardian Date

- -

Part 2
DEMOGRAPHICS FORM

If you decide that you would like your child to participate in this study, please complete the demographic questions below about your child so the findings from this study can be used to understand other children.

Child's name: ____
First name Last name

Child's date of birth: ________________
Month/ Day/ Year

Child's gender (circle one): male female

Is your child of Hispanic descent? (circle one): yes no

Which racial background best describes your child? (check all that apply):

□ White/Caucasian

□ Black/African American

□ Asian/Pacific Islander

□ Other

Appendix D

Previous Testing and Scores

Student's Name	WISC-IV FSIQ	RIAS CIX	WJ III COG GIA	TCAP

Appendix E
ASSENT FOR STUDENTS

Your mother, father, or guardian has told me that you can work with me today, but I need your permission, too. I need to make sure that you know about what we'll do together and that you want to work with us. I think it will be fun.

- I am asking you do some exercises that are like assignments you already do in your classroom. You will be asked to answer some questions and solve some problems. Today we will work together for about 90 minutes.

- You can decide at any time that you don't want to do the thinking tasks, and it will be OK. If you have questions about these exercises, you can ask me at any time.

- I want you to do your best, but you do NOT get a grade on these exercises. They are just for us to learn about how kids think and solve problems. I won't tell your teacher, your friends, or anyone else at the school how you did on these exercises. Also, anything you tell me today will not be shared with your teacher, your friends, or your parents unless someone could be harmed.

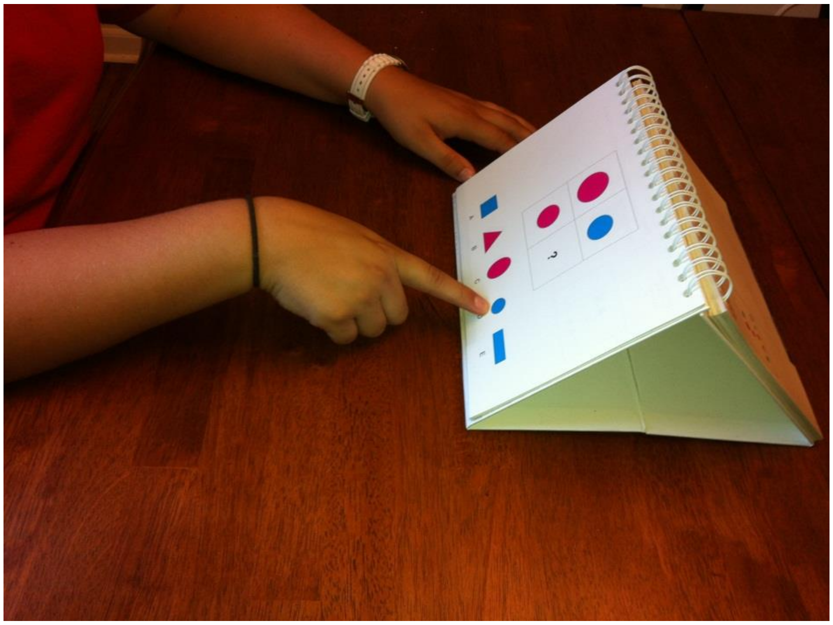
- As a thank you for working with us you will receive a small prize such as a pencil, sticker, or eraser.

If you want to work with us, write your name your name in this box.

□

____________________________________ ______________________________
Examiner's Signature Date

Appendix F
Examples of Video Recording



Appendix G

Examiner Demographics 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: __________

Number of psychoeducational assessment reports written in practice: __________