ERRORS ON COGNITIVE ASSESSMENTS ADMINISTERED BY GRADUATE STUDENTS AND PRACTICING SCHOOL PSYCHOLOGISTS

by

Erika Rodger

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Psychology

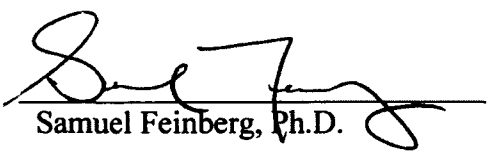
Fairleigh Dickinson University

2011

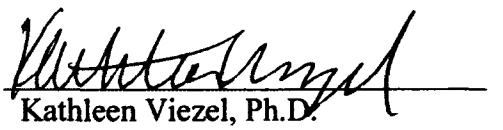
Approved by:

Ronald Dumont, Ed.D

Chairperson of Supervisory Committee



Samuel Feinberg, Ph.D.



Kathleen Viezel, Ph.D.

College Authorized to Offer Degree:

University College: Arts • Sciences • Professional Studies

Date: October 19, 2011

Dedication

This is dedicated to my parents who were there for me from the beginning, and who have supported and encouraged me throughout everything. To Jon, who was very patient with me during this process and stood by me. To both my grandmothers, who provided the foundation of my character.

Abstract

Errors on Cognitive Assessments Administered by Graduate Students

and Practicing School Psychologists

by Erika Rodger

Chairperson of the Supervisory Committee:

Professor Ron Dumont

Cognitive assessments are prevalent in U.S. history and policy, and are still very widely used for a variety of purposes. Individuals are trained on the administration and interpretation of these assessments, and upon completion of a program it should be assumed that they are able to complete an assessment without making administrative, scoring, or recording errors. However, an examination of assessment protocols completed by students as well as practicing school psychologists reveals that errors are the norm, not the exception. The purpose of this study was to examine errors committed by both master's and doctoral-level students on three series of cognitive assessments as well as errors made by practicing school psychologists. Three primary measures of intellectual assessment, the Wechsler scales, the Woodcock Johnson III Tests of Cognitive Abilities (WJ III COG), and the Differential Ability Scales—Second Edition (DAS-II) were examined. Errors were tabulated and analyzed from 295 Wechsler protocols, 257 DAS-II protocols, and 258 WJ III COG protocols completed by master's and doctoral-level students and practicing school psychologists. Errors were delineated by type and source, and also analyzed for identifiable themes. Results of the study revealed that most errors are manualized, in that

what an examiner recorded directly contradicted an instruction given in a test manual. The most common errors found were failure to record responses verbatim, incorrect calculation of raw score, and failure to administer sample items. Based on the outcome of this study, it is recommended that programs instructing students on how to administer cognitive assessment provide ample feedback, and it is recommended that practicing psychologists maintain best practices and take part in continuing education regarding cognitive assessments.

Table of Contents

List of Tables

Acknowledgments

Thanks to my committee chair, Dr. Ronald Dumont, who was very understanding and motivating during this long journey. He encouraged my progress and made sure I was on track. Thanks to the other members of my committee Dr. Samuel Feinberg and Dr. Kathleen Viezel for their support, feedback, and advice during this process.

Chapter 1

Introduction

Under certain circumstances, cognitive assessments are required by U.S. law and policy (Social Security Administration, 2008; 18 U.S.C. § 3596, 2007; U.S. Department of Defense, 2005). These assessments are essential for understanding areas of strength and weakness in individuals and for providing data in legal and policy decisions. The results of these assessments have been used in the determination of learning disabilities, giftedness, qualification for service in the armed forces, social security eligibility, employment opportunities, and capital punishment sentences. Therefore, it is clear that accurate results from cognitive assessments is crucial, and training programs must be equipped to ensure error-free administration.

History of Cognitive Assessment in United States Policy

The administration of a cognitive assessment to determine a general mental ability score, sometimes referred to as an IQ, has many uses within the educational, legal, and military systems of the United States. U.S. history includes several instances where the determination of an individual's IQ determined aspects of life such as marriage and sterilization. From the late 19th century to the mid-20th century, sterilization of the "feeble-minded" was legal and encouraged (Miklos & Carlson, 2000, p. 155). Connecticut enacted a marriage law in 1896 prohibiting marriage for the "epileptic, imbecile, or feeble-minded" (Ellis, Abrams, & Abrams, 2009, p. 391). Indiana passed an involuntary-sterilization law in 1907 intended to "prevent procreation of confirmed criminals, idiots, imbeciles, and rapists" (Stern, 2007, p. 9). These policies have fallen out of favor and are no longer implemented. However, the results of cognitive assessments continue to have implications in many areas of modern U.S. policy. One area where

cognitive assessment is regularly utilized is in the U.S. school system, particularly in the classification of learning disability.

Classification of Learning Disability

In the past, it was common for schools to utilize an IQ/achievement discrepancy model to determine whether a student had a learning disability (Fletcher, 1992). The IQ/achievement discrepancy model stipulated that a "severe discrepancy" (Fletcher, 1992, p. 546) existed between several elements, including age-equivalent score, chronological age, IQ, and level of achievement. While there has been a national push toward the response-to-intervention (RTI) model, the IQ/achievement discrepancy model is still used in many schools (Aaron, Joshi, Gooden, & Bentum, 2008). Recent discussions on the National Association of School Psychologists Listserv, as well as anecdotal evidence, reveal that learning disability identification through severe discrepancy is still an issue under much debate. Federal law allows schools the option to continue to identify students for special education as learning disabled through the use of severe discrepancy analysis, and with this option, it stands to reason that schools will continue to do so. There are other areas in the school system where a cognitive assessment may be used. For example, cognitive assessments in schools not only can indicate why a student is not achieving to his or her ability but also can be used to determine whether a student requires a gifted program to be intellectually challenged in the classroom. Thus, cognitive assessments can also be used to determine giftedness.

Determination of Giftedness

There is no universal definition of giftedness, and prerequisites for admission to a gifted program vary depending on the particular program. One definition (Johnsen, 2004) stated that "gifted and talented" means "students, children or youth who give evidence of high performance

capability in areas such as intellectual, creative, artistic, or leadership capacity" (p. 2). Other definitions rely on an IQ score entirely. Gross (2004) revealed that there are different classifications of giftedness ranging from *mildly gifted* (starting at an IQ score of 115) to *profoundly gifted* (with an IQ score of at least 180). Giftedness determination and education is not regulated by federal law, although there are some recommendations put forth by the U.S. Department of Education on a state-to-state basis. For example, the New Jersey State Board of Education (2005) defined gifted and talented as "those students who possess or demonstrate high levels of ability, in one or more content areas, when compared to their chronological peers in the local district and who require modification of their educational program if they are to achieve in accordance with their capabilities" (p. 10). Many gifted and talented programs rely on IQ testing as a part of the admissions process, with some programs taking the top 2% of the normative population, and others taking the top 5% of the normative population. Several sources (Clark, 2002; Johnsen, 2004; Sternberg, 2004) indicated that several types of assessments should be used to determine giftedness, with a cognitive assessment being one of them.

Not only are cognitive assessments utilized with regularity in school systems, but they are also utilized in other areas of U.S. policy. For example, they are utilized in the determination of qualification for military service in the United States.

Qualification for Military Service

Part of the recruitment process for the U.S. Military is to take the Armed Services Vocational Aptitude Battery (ASVAB; U.S. Department of Defense, 2005). The ASVAB was initially developed in 1968 and has been utilized by the military since 1976 for the purpose of identifying individuals appropriate for the military as well as possible branches of the military the individual may serve. The ASVAB contains nine subtests including Word Knowledge,

Arithmetic Reasoning, Paragraph Comprehension, and Mathematics Knowledge. The overall performance on the entire ASVAB determines possible qualification for military occupational specialties. An individual's performance on the previously mentioned four subtests forms the Armed Forces Qualifying Test (AFQT) score. This score determines whether an individual is qualified to enter the U.S. military. The AFQT has been found to be an indicator of premorbid intelligence when compared with the Multidimensional Aptitude Battery (Orme, Brehm, & Ree, 2001). The AFQT score is reported as percentile ranks between 1 and 99, with an individual's percentile score indicating where they fell within the reference group. U.S. Department of Defense (2005) policy requires an individual to score at least at the 10th percentile on the AFQT, citing difficulty in training individuals who fall at or below the 9th percentile. The policy further notes that only 4% of recruits may score within the 10th to 30th percentile. The Army requires a minimum AFQT score of 31, the Navy requires a minimum score of 35, and the Coast Guard requires a minimum score of 45. The U.S. government utilizes cognitive assessment not only to determine placement in military service, but also for determining eligibility for social security disability benefits.

Determining Social Security Disability Benefits

The U.S. Social Security Administration provides social welfare and social insurance to a wide range of individuals, including the disabled. One of the possible disabilities that can qualify for benefits is mental retardation. The requirements that determine benefits for the disability of mental retardation include "a valid verbal, performance, or full scale IQ of 59 or less" (Social Security Administration, 2008, p. 53). Additionally, the requirements also stipulate that one may qualify for the disability with "a valid verbal, performance, or full scale IQ of 60 through 70 and a physical or other mental impairment imposing an additional and significant

work-related limitation of function" (Social Security Administration, 2008, p. 53). A valid cognitive assessment needs to be administered by a trained professional in order to ascertain these requirements for benefits.

Cognitive Ability in Employment

IQ testing as a prerequisite to employment is frowned upon due to the "substantial adverse impact on employment opportunities for members of several racial and ethnic minority groups" (Murphy, 2002, p. 173). The U.S. Civil Rights Act as interpreted in the 1971 Supreme Court decision in *Griggs v. Duke Power Co.* (1971) bars companies from using cognitive assessments as a deciding factor in hiring employees. Tests are only permitted to be used alongside a subjective hiring process and must only be used when the test is reasonably related to the job for which the test is required (*Griggs v. Duke Power Co.*, 1971). This proves to be a hurdle for companies, as cognitive ability tests have been found to be the "best single predictor of job performance" (Murphy, 2002, p. 185). However, employers now have the legal burden to prove the business necessity of the test. Certain companies offering employment opportunities utilize tests that correlate with IQ tests. Microsoft, for example, provides verbal brain teasers that determine abstract thinking skills without being a standardized cognitive assessment (Karlgaard, 2005). This test is used in conjunction with the interview process and is one method many technology and consulting companies, such as Google, Amazon, McKinsey, and Accenture (Kaplan, 2007), are beginning to use to assess reasoning skills when determining employee selection. Another employment opportunity that uses a cognitive assessment is the National Football League. The National Football League utilizes the Wonderlic Cognitive Ability Test to assess the aptitude of predraft prospective players (Kuzmits & Adams, 2008).

The National Football League administers the 12-minute test to individuals and utilizes the score for determining the appropriate position for a player (Super, 2006).

Cognitive Assessment in Capital Punishment

The U.S. Supreme court ruled in 2002 in the case of *Atkins v. Virginia* (2002) that mentally retarded criminals may not be executed on the grounds that it would constitute cruel and unusual punishment, prohibited by the Eighth Amendment. In this particular case, Atkins was administered a standardized cognitive assessment, and his full-scale IQ was determined to be 59. The case further noted that his score would have "automatically qualified him for Social Security benefits" (*Atkins v. Virginia*, 2002) and it was determined that he was mentally retarded. His death penalty was overturned due to this determination, and the case changed the requirements for a sentence of capital punishment within the United States. Federal law now prohibits the execution of offenders who are mentally retarded, and offenders who are sentenced to death are executed pursuant to the laws of the state in which the death penalty was imposed, as stated in the Death Sentence section of the U.S. Code (2007). Eighteen of the 38 states that permit the death penalty prohibited the execution of mentally retarded prisoners even before the Atkins ruling. At least 10 additional states have prohibited the execution of mentally retarded prisoners since the Atkins ruling (DeMatteo, Marczyk, & Pich, 2007). Individual states have their own criteria regarding the definition of mental retardation as it relates to capital punishment. For example, the state of Arkansas requires an IQ of under 65, and the state of Illinois requires an IQ of under 75. Several other states (e.g., Maryland and Rhode Island) require the IQ to be two standard deviations below the mean. There is no federally mandated definition for mental retardation in criminal matters. Given the impact of the Atkins ruling and the protection from the death penalty of criminals with a measured IQ under 65 to 75 depending

on the state, it is clearly important to have qualified trained professionals who are able to administer and interpret standardized cognitive assessments and who are able to do them correctly. Research shows scoring errors or administrative errors can significantly change the overall IQ (Alfonso, Johnson, Patinella, & Rader, 1988; Loe, Kadlubek, & Marks, 2007; Slate & Jones, 1993), and errors made may be the difference between life and death for these individuals.

Implications for Practice

Due to the use of IQ testing within U.S. law and policy, it is vital that the results of a cognitive assessment administered by a trained professional be correct. Research has shown that graduate students learning to administer assessments make mistakes throughout the learning process (Alfonso et al., 1998; Loe et al., 2007; Patterson, Slate, Jones, & Steger, 1995; Slate & Jones, 1990). Research has further shown that certified and trained professionals continue to make scoring errors even while they are in practice (Brazelton, Jackson, Buckhalt, Shapiro, & Byrd, 2003; Slate & Jones, 1993; Slate, Jones, Coulter, & Covert, 1992). It is clear that attention must be given to administration and scoring errors on cognitive assessments.

One aspect of graduate-level training for school psychology requires learners to administer cognitive assessments (Alfonso, LaRocca, Oakland, & Spanakos, 2000; Belter & Piotrowski, 2001; Cody & Prieto, 2000; Oakland & Zimmerman, 1986). Learning to administer cognitive assessments typically includes instruction on several types of assessments, such as the Wechsler scales, the WJ III COG, and the DAS-II, and the requirement of administering a certain number of each test. It is expected that graduate students will make administrative errors during their training, and learning how to avoid these errors is important in learning the assessment. Because administration and scoring errors on these assessments may cause different IQ scores than the tested individual should have achieved (Alfonso et al., 1998; Belk, LoBello, Ray, &

Zachar, 2002; Patterson et al., 1995; Slate & Jones, 1993), it is vital to examine the types of errors that are commonly made, the types of errors that persist throughout training, and the difference between graduate students and practicing school psychologists. Increased knowledge in this area will help drive the future training of school psychologists and affect the current administrative methods of practicing school psychologists.

Practicing school psychologists should also be aware of the common types and frequencies of errors, as becoming certified does not guarantee perfection. Practitioners utilizing cognitive measurements should be aware of the types of errors found even in experienced testers. Errors on a protocol can have potentially disastrous effects (e.g., in the determination of capital punishment) for the individuals tested. Practitioners should thus be concerned with errors on a protocol to "strive to benefit those with whom they work and take care to do no harm," as cited in the Ethical Principles of Psychologists and Code of Conduct of the American Psychological Association (APA, 2010, p. 3). Additionally, should a given case become a legal battle, one's credibility would be severely damaged if administrative errors were discovered.

This dissertation examined the errors that graduate students and practicing school psychologists make on cognitive assessments. Three primary measures of intellectual assessment, the Wechsler scales (Wechsler, 2002, 2003a, 2008), the WJ III COG (Woodcock et al., 2001a), and the DAS-II (Elliott, 2007), were examined. The Wechsler series is arguably the most commonly taught cognitive assessment in graduate programs (Alfonso, Oakland, LaRocca & Spanakos, 2000; Belter & Piotrowski, 2001; Cody & Prieto, 2000; Oakland & Zimmerman, 1986). The Wechsler series has been in existence since 1949 and contains a preschool test, the Wechsler Preschool and Primary Scale of Intelligence—Third Edition (WPPSI-III; Wechsler, 2002), most recently revised in 2002; a school-age child test, the Wechsler Intelligence Scale for

Children—Fourth Edition (WISC-IV; Wechsler, 2003a), most recently revised in 2003; and an adult test, the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV; Wechsler, 2008), most recently revised in 2008. The WJ III COG (Woodcock et al., 2001a) is commonly taught and is often administered when determining a learning disability (Barnes, Fletcher, & Fuchs, 2007; Cody & Prieto, 2000). It has been in use since 1977 and was most recently revised in 2001. The DAS-II is increasing in popularity and is now being commonly taught in many schools, as demonstrated by responses to a question posed to the Trainers of School Psychologists Listserv (Rodger, 2011). In the Listserv thread, 32 respondents indicated that the DAS-II is formally taught at 23 programs, four programs make the DAS-II available in an internship or practicum, and five programs do not offer the DAS-II. The DAS-II has been used since 1990 and was most recently revised in 2007. These measures are taught in graduate programs and are utilized in schools, and students who are learning the assessments and professionals who are trained to give the assessments continue to make administrative and scoring errors on them. This is a pervasive problem that warrants continued investigation on how graduate programs and professional development programs may better instruct and support those who are administering these measures of intellectual assessment to ensure accurate assessment results.

Typical Administrative Errors

Several studies have been conducted that examine the frequency and types of administrative errors on the Wechsler scales. They include examining the administered protocols of graduate students (Alfonso et al., 1998; Belk et al., 2002; Conner & Woodall, 1983; Loe et al., 2007; Patterson et al., 1995; Slate & Jones, 1990; Warren & Brown, 1972), examining scoring errors on provided protocols (Brazelton et al., 2003; Erdodi, Richard, & Hopwood,

2009), and examining errors made by practicing school psychologists (Brazelton et al., 2003; Erdodi et al., 2009; Slate & Jones, 1993; Slate et al., 1992). There are differing error-recording systems in every study. For example, some studies examine whether errors happen once per subtest, whereas others record every instance of error, thus increasing the number of errors.

Administrative Errors on Record Forms Completed by Graduate Students

Warren and Brown (1972) examined individual subtests from 120 WISC and 120 Stanford–Binet Intelligence Scale protocols completed by 40 graduate students. Of the 1,873 subtests examined, 1,939 total errors were found. Examiner errors accounted for a change in Full Scale IQ (FSIQ) scores on 50 WISC and 39 Binet protocols. The most common WISC error was the failure to record a response, and the most common Binet error was failure to follow procedures delineated in a manual, such as determining final scoring during test administration, not afterwards. Warren and Brown also found that class discussion and feedback to the students did not affect the number of errors committed, indicating that examiners did not improve with practice with these methods of instruction.

Conner and Woodall (1983) examined the effects of experience and structured feedback on WISC-R error rates. Ten graduate students submitted 150 protocols. Each was evaluated for different types of errors on response error, IQ reporting error, administrative error, and mathematical error. Results indicated that scoring errors remained despite experience testing, and only total errors and administrative errors significantly decreased with a structured feedback instrument. Response scoring errors accounted for 58% of the errors on each protocol, and it was postulated that as examiners gain experience, they rely on memory to score a response rather than utilizing the manual.

Slate and Jones (1990) examined WISC-R protocols of 26 graduate students learning to administer the assessment. The graduate students found that all 216 protocols examined contained at least one error, and that, on average, there were about 11.3 errors per protocol. The most frequent error was the failure to record responses and incorrect point assignment. Students were also found to be more likely to give too many points to a response rather than too few. Other errors included failure to query, basal/ceiling problems, incorrect raw score calculations, and incorrect age calculations. The study further revealed that student errors did not decrease after five administrations, and only decreased slightly after 10 administrations. It was recommended that instructors focus on difficult-to-score subtests and that students be encouraged to check over protocols. Another study examined errors on the WAIS-R.

Patterson et al. (1995) studied graduate student administration of the WAIS-R. Of 149 protocols from 22 graduate students examined, the average number of errors was 41.2. There was no improvement noted over five administrations, and no significant improvement after 10 administrations. Failure to record responses was the most common error, followed by assigning too much credit to a response. In these protocols, 71% of the students overestimated the FSIQ, while only 7% underestimated the FSIQ. It was recommended that students be given a list of commonly committed errors and receive immediate feedback.

Alfonso et al. (1998) examined 60 WISC-III protocols from 15 graduate students to determine the frequency and types of administration errors. A total of 468 errors were found with an average of 31.2 errors per protocol. Errors decreased as the individual gained experience with administering tests when accompanied by class instruction, peer training, and verbal and written feedback. The average number of errors decreased from 14.4 to 5.4 across four protocols. The researchers found that the errors were not consistently associated with a specific

subtest, and failure to query, failure to record responses, adding subtest scores incorrectly, and incorrect IQ scores were the most common errors committed. Alfonso et al. indicated that the Comprehension subtest had the highest percentage of errors per subtest across all 60 protocols, at 11%. The next highest percentage was in the Similarities subtest at 6%. Additionally, the Vocabulary subtest had a similar number of errors when compared to nonverbal subtests. This is contrary to many other studies (Brazelton et al., 2003; Slate & Jones, 1993; Slate et al., 1992) that found all Verbal Scale subtests had problems with scoring.

Belk et al. (2002) examined WISC-III administrative errors. Twenty-one graduate students submitted 100 WISC-III protocols for evaluation. All protocols contained errors, with the most common error being failure to record responses. There was a mean of 45.2 errors per protocol. Overall, the impact of these scoring errors resulted in an overestimation of FSIQ on 46% of the protocols and an underestimation on 21% of the protocols. The average change in FSIQ was 0.83 points. Students were found to be more likely to assign too many points to a response than too few, as in the 1990 Slate and Jones study. There was no improvement in test administration when comparing first administration to last administration.

Loe et al. (2007) examined 51 WISC-IV protocols administered by 17 graduate students-in-training to determine the frequency of examiner errors and the impact of those errors on the Index and FSIQ scores. Errors were divided into three categories: administration, computation, and recording. They found that students committed errors on 98% of the protocols, and averaged 25.8 errors per protocol. The most common errors included failure to query, assigning too many points to an answer, and failure to record responses. The researchers found that administration errors were the most common, and specifically that individuals failed to query, failed to record responses, and assigned too many points to a response. When administration and computation

errors were corrected, 35% of all FSIQ and Index scores were altered. Scores often dropped once corrected, indicating that abilities had been overestimated. The researchers indicated that due to the continued high frequency of query and scoring errors, the WISC-IV remains difficult for novice examiners to learn despite test revisions and improvements in scoring and querying criteria, specifically that "mistake free student protocols are as difficult to find on the WISC-IV as with previous versions of the test" (Loe et al., 2007, p. 244). They noted that failure to record responses was the most common error, and that the Block Design subtest was the primary source for the error of failure to record examinee responses. They also found that graduate students did not make significantly fewer errors after three practice administrations.

Several studies indicated that repeated administrations decreased the number of errors committed. Conner and Woodall (1983) found that errors dramatically decreased on the WISC-R when examiners were given feedback about type and number of errors committed. Alfonso et al. (1998) also found that when class instruction, peer training, and verbal and written feedback were provided, the number of errors decreased per protocol as students gained experience.

However, most studies support the notion that graduate students continue to commit errors on cognitive assessments despite repeated practice administrations. Warren and Brown (1972) found that there was no significant improvement with feedback, and Slate and Jones (1993) found that errors did not decrease over five administrations and only slightly decreased over 10 administrations. Belk et al. (2002) found that no improvement occurred over several administrations, and Loe et al. (2007) found that error rates did not improve over three practice administrations.

Prefilled Unscored Record Forms to Examine Scoring Errors

Brazelton et al. (2003) distributed three WISC-III protocols to individuals of varying levels and fields of education for scoring. The protocols contained responses that had not been scored, and the participants were asked to score the responses. A total of 126 protocols were examined and revealed that Comprehension, Vocabulary, and Similarities were the subtests most often scored incorrectly. The researchers found that none of the 126 protocols were error-free and most protocols had multiple errors. There were errors in 98% of the Comprehension subtests, 96% of the Vocabulary subtests, 75% of the Similarities subtests, 37% of the Picture Completion subtests, and 34% of the Information subtests. In Brazelton et al.'s study, the relationship between experience and scoring errors was stronger (individuals with more practice made fewer errors) than the level of education or position and scoring errors. The researchers indicated that this implies that "the adage, 'practice makes perfect,' may, in fact, apply to scoring of the WISC-III" (p. 7).

Erdodi et al. (2009) examined just the Vocabulary subtest of the WISC-IV. Three fictitious unscored Vocabulary subtests were given to graduate students. The protocols were created so that if scored correctly, scaled scores of 4, 10, and 16 would be obtained, and the study examined how closely the graduate students came to the intended score. Forty-six graduate students in two clinical psychology programs were given the protocols to score. The researchers found that "graduate students were more prone to make scoring errors in the extremely low and superior ranges of the IQ distribution" (p. 383). They also noted that some of the errors in scoring resulted in clinically meaningful deviation from the true score that should have been obtained. Further results also found that master's level students made more errors than doctoral-level students.

Clinical Experience and Scoring or Administrative Errors

Slate et al. (1992) conducted a study utilizing a sample of practicing school psychologists. They examined 56 WISC-R protocols completed by one certified and eight licensed practitioners. The purpose was to examine whether administrative errors were a function of the learning process or whether mistakes continue into actual practice. Errors were found on all protocols, with an average number of 38.4 errors. The most problematic subtests were found to be Picture Completion, Picture Arrangement, and Vocabulary. All the practitioners were found to have failed to record at least one response. It was also found that it was more common to give too many points than too few.

Slate and Jones conducted a similar study in 1993 utilizing the WAIS-R. Fifty protocols from eight experienced practitioners were examined. Errors were found on all 50 protocols, with an average number of 36 errors per protocol. The number of errors ranged from 13 to 103. The most problematic subtests were Vocabulary, Comprehension, and Similarities. Practitioners were more likely to give too many points to a response than too few. After the practitioner errors were corrected, the FSIQ was changed on 26 of the protocols. Of these corrected FSIQs, when an FSIQ did change, 88% of the time it resulted in a lower FSIQ than those assigned by the practitioner, and 12% of the time it was higher. All the changed IQs were within 5 points of the originally calculated FSIQ. The researchers indicated that, consistent with previous research, practice does not make perfect, and in spite of professional experience, practitioners are more prone to make errors than graduate students based on previous research.

Brazelton et al. (2003) provided surveys and WISC-III protocols to 124 practitioners to determine the relationship between highest degree, number of WISC-III protocols administered, and current scoring errors on protocols. The surveys provided information regarding current

position, field of highest degree earned, highest degree earned, and number of WISC-IIIs administered during one's career. The protocols used in this study had actual student responses but no scores. Of the professionals surveyed, 68% held the position of school psychologist, 16% were employed in private practice, and 12% were interns. Forty-three percent of the respondents listed school psychology as their field of highest degree, 24% listed clinical psychology, 11% listed special education, and 10% listed counseling. In response to highest degree attained, 41% earned master's + hours for certification, 22% earned an Ed.S., 20% had earned a doctorate, and 14% had earned a master's. For the number of WISC-III protocols administered in their career, 58% reported over 200 protocols, 18% reported 101-200 protocols, 5% reported 51-100 protocols, 13% reported 11-50 protocols, 6% reported 1-10 protocols, and 1% reported 0 protocols. The relationship between number of scoring errors and experience was significant when experience was defined as number of protocols administered. Those who reported administering more than 100 WISC-III protocols made fewer errors than those who had administered 10 or fewer. There was also a relationship between scoring errors and position, as those working in schools as school psychologists or psychometrists made fewer errors than those in other positions.

Erdodi et al. (2009) conducted a similar study and examined adherence to manualized scoring of WISC-IV protocols. Participants were given fictitious WISC-IV Vocabulary subtest responses and asked to score each item. The subtests were constructed to yield scaled scores of 4, 10, and 16. Participants committed significantly more scoring errors in the extremely low and superior IQ distribution. Those with more clinical experience made more errors, and the researchers indicated that this implied the practitioners used memory to score a test rather than

the manual as they gain more clinical experience. Familiarity with administration and scoring of protocols resulted in more errors, whereas adherence to the manual resulted in fewer errors.

Table 1 summarizes several studies conducted examining errors on the Wechsler scales. All studies found errors in administration of cognitive assessments regardless of level of education, and almost all protocols examined across all studies had at least one error. The most common error was failure to record errors verbatim, and on many of the studies the errors were found to affect the FSIQs. Several of these studies found no improvement in examiner error over time.

Non-Wechsler Tests

Ramos, Alfonso, and Schermerhorn (2009) examined administration and scoring errors committed by graduate students. On the 108 WJ III COG protocols from 36 graduate students examined, 500 errors were found. The mean errors per test were found to be approximately 4.63. Ramos et al. noted that "it is important to note that 167 of the total errors (33%) were made on 5 test records, whereas 50 test records (46%) had either 0 errors or 1 error" (p. 654). The most problematic subtests in terms of number of errors were Verbal Comprehension, Visual-Auditory Learning, General Information, and Retrieval Fluency. The most common errors were found to be incorrect ceilings, failure to record examinee response, and failure to circle the correct row for the total number incorrect.

Table 1

Summary of Results of Previous Studies Examining Administrative and Scoring Errors on the Wechsler Scales

Study	Sample	Instrument investigated	Major findings
Warren & Brown (1972)	40 graduate students	WISC, Binet	Failure to record response was the most common WISC error, failure to follow procedure the most common in Binet; errors impacted the FSIQ on 50 WISC protocols
Sherrets, Gard, & Langner (1979)[a]	39 psychologists, interns, practicum students, school psychologists, and psychometricians	WISC	89% of examiners made at least one error; most common errors were in addition of scores
Conner & Woodall (1983)	10 graduate students	WISC-R	Experience with administration resulted in decreased errors
Slate & Chick (1989)[a]	14 graduate students	WISC-R	All subtests were found to have some error; errors on 66% of the protocols resulted in changes to the FSIQ
Slate & Jones (1990)[a]	26 graduate students	WISC-R	All protocols contained error; an average of 11.3 errors per protocol; frequent errors included failure to record examinee responses, incorrect point assignment, and inappropriate questioning
Slate, Jones, Coulter, & Covert (1992)[a]	9 certified psychological examiners	WISC-R	All protocols contained error; an average of 38.4 errors per protocol including failure to record response; errors on 81% of the protocols resulted in changes to FSIQ
Slate & Jones (1993)	8 practitioners with extensive experience	WAIS-R	All protocols contained error; an average of 36.9 errors per protocol including failure to record response; practitioner error on 27 of 50 protocols changed FSIQ
Patterson, Slate, Jones & Steger (1995)	22 graduate students	WAIS-R	Most common errors were failure to record response, too few or too many points, and failure to query; no significant improvement was noted after 10 administrations
Alfonso, Johnson, Patinella, & Rader (1998)[a]	15 graduate students	WISC-III	An average of 7.8 errors per protocol; frequent errors included failure to query, failure to record responses verbatim, reporting incorrect FSIQs, reporting incorrect VIQs, and incorrect addition of scores
Belk, LoBello, Ray, & Zachar (2002)	21 graduate students	WISC-III	All protocols had errors, average of 45.2 errors per protocol; most common error was failure to record response, most problematic subtests were Vocabulary, Comprehension, and Similarities
Loe, Kadlubek, & Marks (2007)[a]	17 graduate students	WISC-IV	An average of 25.8 errors per protocol; common errors were failure to query, assigning too many points to a response, failure to record an examinee's response, and inaccurate test composite scores, resulting in incorrect FSIQ and Verbal Comprehension Index
Erdodi, Richard, & Hopwood (2009)	46 graduate students	WISC-IV	On the Vocabulary subtest, graduate students were more prone to scoring errors in the extremely low and superior ranges of the IQ distribution

Note. WISC = Wechsler Intelligence Scale for Children; WISC-R = Wechsler Intelligence Scale for Children—Revised; WISC-III = Wechsler Intelligence Scale for Children, Third Edition; WISC-IV = Wechsler Intelligence Scale for Children, Fourth Edition; WAIS-R = Wechsler Adult Intelligence Scale—Revised; FSIQ = Full Scale IQ; VIQ = Verbal IQ.

[a]Table format from Ramos, Alfonso, and Schermerhorn (2009).

An extensive database search was conducted using key terms such as *differential ability scales*, *DAS-II*, *DAS*, *errors*, *cognitive*, *cognitive assessments*, and *Elliott*, but no research regarding the Differential Ability Scales was found. Databases utilized included Academic Search Premier, E-Journals, ERIC, Primary Search, PsycARTICLES, and PsycINFO.

Recommendations to Reduce Examiner Error

Examiner error is obviously an area of concern, and some studies have attempted to provide recommendations to graduate trainers that may lower the number of errors and make the task of teaching assessment easier. Egan, McCabe, Semenchuk, and Butler (2003) examined the difference in errors committed on the Wechsler series by graduate students under two conditions. Twenty individuals enrolled in a graduate-level cognitive assessment course were broken into an experimental and a control group. Those in the control condition received a lecture and demonstrations on proper administration and scoring of assessments. Those in the experimental condition had the same lecture and demonstration as the control group, but they were also required to maintain a portfolio of completed protocols that were reviewed before every subsequent administration. Individuals maintaining a portfolio of protocols made significantly fewer mistakes as the semester progressed when compared to the control group, who did not keep a portfolio.

How Graduate Schools Teach Cognitive Exams

Methods of teaching cognitive assessments differ among graduate programs. The type of cognitive assessment, the number of required assessments, the materials available for use, even the time commitment to the course differs from program to program. Several surveys have been conducted to examine the commonalities and differences between graduate programs offering coursework on cognitive assessment.

A national survey conducted in 1986 by Oakland and Zimmerman obtained individual mental assessment course information from 49 instructors in the United States. The survey was intended to examine the typical assessments covered by the course, the average time commitment for the course, the number of protocols administered, qualifications of instructors, how many credit hours were offered, how many semesters were necessary for completion of the course, minimum and maximum number of students for the course, and other specifics of the course. They found that 94% of programs utilized the WISC-R, 88% utilized the WAIS-R, 86% utilized the Stanford-Binet, and 82% utilized the WPPSI. When the WISC-R was utilized, students were required to administer on average seven protocols, seven practice tests, and three observations. Seventy percent of the programs utilized the skills of a graduate teaching assistant and 56% of the programs required students to find their own individuals to assess. In 6% of the programs, the responsibility of finding volunteers fell on the instructor, and in 17% of the programs the students and instructor jointly found volunteers. Students reported an average of 11 hours per week devoted to the course.

In a follow-up survey conducted by Alfonso et al. (2000), 97 instructors of graduate-level courses on individual cognitive assessment were given questionnaires about common practices in their classes. The questionnaire was a revised version of the Oakland and Zimmerman version distributed in 1986. They found that 94% of courses taught the WISC-III, 80% taught the WAIS-R, 64% taught the WPPSI-R, and 74% taught the Stanford-Binet Intelligence Test-Fourth Edition. Seventy-six percent of the instructors used graduate assistants to assist in grading protocols. In courses that required the WISC-R, the average number of protocols, reports, and competency exams required were five, four, and one, respectively. On average, instructors spent

4 to 5 hours per week grading protocols, and graduate students spent 2 hours per week grading protocols.

In a national survey, Cody and Prieto (2000) surveyed 94 APA-accredited graduate programs in the United States and Canada to determine how intelligence testing courses are taught at the graduate level. A survey was distributed that mirrored the previous Oakland and Zimmerman survey. Cody and Prieto found that 77% of programs gave students access to two-way mirrors and videotapes, and 96% of the students were permitted to borrow the testing kits from the graduate school.

Forty-two percent of the classes were required to find their own volunteers. Cody and Prieto (2000) also found that 92% of the programs utilized at least one graduate teaching assistant, which was a dramatic increase from the 1986 study. Among the tests listed in the survey, the WAIS-R or WAIS-III was taught in 91% of the programs, and the WISC-R or WISC-III was taught in 75% of the programs. The next most frequently taught instruments included the Woodcock–Johnson Psychoeducational Battery (22%) and the Kaufman Assessment Battery for Children (17%).

In a survey directed to APA-approved doctoral programs in clinical psychology, Belter and Piotrowski (2001) asked questions regarding the types of assessments taught and whether emphasis has increased, decreased, or stayed the same for these assessments. Eighty-two percent of those surveyed indicated that intelligence testing was required within the program; the primary intelligence tests utilized were the WAIS (94%), WISC (74%), Stanford-Binet (27%), and Kaufman scales (17%). The survey results also reported that 59% of the programs found the WAIS-R/WAIS-III to be essential for practice. IQ tests were rated as increasingly important in

7% of the programs, stayed the same in 86% of the programs, and decreased in importance in 6% of the programs.

An examination of these studies clearly indicates that the Wechsler series is the most commonly taught assessment. Even in clinical psychology doctoral programs, the Wechsler series is taught when IQ testing is a requirement.

Overall, it is obvious that IQ testing is crucial in many areas of U.S. law and policy, and the accuracy of the results must be reliable. IQ testing is a requirement of many school psychology and clinical psychology programs, with all programs surveyed requiring students to administer several types of assessments. Resulting research showed that graduate students make a large variety of mistakes, and most continue to make mistakes even through the end of the program. Practicing school psychologists who have been through the certification process were also shown to consistently make scoring errors. Given how crucial it is for professionals to obtain accurate IQ scores, it is very important to examine what types of errors on cognitive assessments continue despite test revisions. The present study examined errors committed by both master's and doctoral-level students on three series of cognitive assessments and errors made by practicing school psychologists. Thus far, there have been no studies that included an examination of errors across tests, and there have been no studies on the DAS-II. This study provided data regarding current trends in administrative and scoring errors on different types of cognitive assessments throughout levels of experience.

Hypotheses

1. None of the given tests (Wechsler series, DAS-II, WJ III COG) will exhibit more total administrative errors than the others.

2. There will be no difference between the average number of errors committed by master's students versus doctoral students.
3. Practicing school psychologists will make the same average number of Wechsler series errors as students learning to administer cognitive assessments.

Chapter 2

Method and Procedures

Research objectives are addressed through statements of methods used and procedures followed. A description of data collection for this archival study is described. Procedures for calculating, scoring, and administration errors also are described.

Data Collection

Data were utilized in the form of cognitive assessment record forms completed as part of cognitive assessment courses at a university in the northeastern United States. One assessment course, taught each fall semester, was for Ph.D. students seeking their degree in clinical psychology. The students did not have any prior experience with cognitive assessments. Another assessment course, taught every spring semester, was for students in a master's/certificate program in school psychology. These students also had no prior experience with cognitive assessments. Approximately three years of data from both courses were analyzed. Both courses met once a week during a 15-week semester, had the same instructor, and had one teaching assistant (TA). The TA's role was to examine protocols handed in by students for administrative, recording, and scoring errors. The TA then gave the corrected protocols to the class professor, who checked the errors found and determined whether they would affect the student's grade. The TA was given training from the class professor on errors.

The classes did not specifically teach the given assessments, but rather required the students to read the manual and practice administering the assessments. As their errors were corrected, the expectation was for the students to learn from their mistakes, and learn not only how to administer the assigned assessments, but also how to read a standardized test manual, so the skills learned could be generalized to other assessments when necessary. A demonstration of

the assessments was not given, though problematic areas and specific questions were addressed in class.

Archival data in an existing database were examined. A requirement of the assessment course was for the students to administer cognitive assessments to others. The cognitive assessments used were the WISC-IV, WAIS-IV, WJ III COG, and DAS-II. Students administered all subtests of each test and computed all scores (e.g., raw, scale, standard) by hand where possible using the procedures and tables in the manuals. The protocols were then handed in and marked for manualized errors by the teaching assistant. The individual errors were recorded onto a spreadsheet. As part of record keeping for the course, records of the errors were kept by the course instructor, and the de-identified database of error information was made available to the researcher. Protocols were de-identified prior to being handed in, with the names of individuals being tested removed, and a code for the professor's reference being used in lieu of student names. The information provided for coursework was recorded in such a manner that participants could not be identified directly or linked to the participants through identifiers. The students were instructed to administer the assessment to individuals who did not display physical, language, or sensory limitations; were not suspected to have an intellectual deficiency; and thus should not require specialized testing considerations.

Additionally, data from a large urban district in the northeast were gathered. These data took the form of protocols administered by practicing school psychologists. The administrations of the cognitive assessments were for educational purposes and intended for educational use, and all identifying information was removed prior to protocol examination. Cognitive assessments were administered as part of eligibility determination for special education. These assessments were not specifically intended for research purposes. Due to de-identification, there was no

chance of the school, the evaluator, or individual personal data being disseminated. Disclosure of the information gained from protocol examination had no chance of being linked to the individual who was assessed. The data requested and then recorded did not include the examiner or the student names, and no identifiers that could link the data to the subjects were recorded. The protocols examined were existing documents where no identifying information was recorded that could link participants or the examiner to the data. There was no chance of placing the individuals at risk of civil or criminal liability or being damaging to the participants' financial standing, employability, or reputation in any instance of data examination as delineated by Title 45 CFR 46.101(b) of the Code of Federal Regulations. This code deals with the protection of human subjects and discusses the proper methods to ensure minimal risk as well as protecting the rights and welfare of human subjects.

Description of Cognitive Assessments Examined

The Wechsler intelligence scales are a series of intelligence assessments designed to assess three different age ranges and are the most commonly taught series in graduate schools (Cody & Prieto, 2000). Protocols from the WISC-IV and WAIS-IV were used in this study.

The WAIS is designed to measure adolescent and adult intelligence between the ages of 16 and 90. The most recent version, the WAIS-IV (Wechsler, 2008), consists of 15 subtests; 10 create a core battery and the remaining five are supplemental subtests. The core subtests are used to compute the FSIQ and four indexes that describe an individual's different functioning levels.

The WAIS-IV has many strengths, especially when compared with its predecessor, the WAIS-III. The measurement of fluid reasoning was enhanced through the addition of the Visual Puzzles and Figure Weights subtests. Working memory measurement was also enhanced with

the addition of Digit Sequencing to the Digit Span subtest. The administration of the WAIS-IV is also fairly easy to learn (Lichtenberger & Kaufman, 2009), and the assessment requires fewer manipulatives than previous WAIS incarnations. All items are dichotomous (score possibility of 0 or 1) with the exception of the Verbal subtests and the Block Design subtest. The average split-half reliability of the WAIS-IV ranged from .97 to .98 and had average reliability coefficients ranging from .96 for Verbal Comprehension to .90 for Processing Speed. There is also strong construct validity for the four-factor structure of the WAIS-IV (Lichtenberger & Kaufman, 2009).

The WISC is designed to measure the intelligence of children and adolescents. The WISC scale is intended for use with children between the ages of 6-0 and 16-11 years. The WISC has been revised four times, with the most recent version being the WISC-IV (Wechsler, 2003a). The WISC-IV consists of 15 subtests divided into core and supplemental subtests. Supplemental subtests can be substituted for core batteries when necessary.

The WISC-IV has several strengths. The basals and ceilings are sufficient to distinguish differing abilities among individuals, and the basals and ceilings are relatively easy to establish. Scoring criteria are relatively more straightforward than the Wechsler predecessors. The protocol provides ample space to write verbatim responses, which enables and strengthens analysis. The WISC-IV yields an FSIQ based on four indexes intended to reveal a general representation of an individual's intellectual functioning. The WISC-IV has excellent internal consistency, averaging .94 for the Verbal Comprehension Index (VCI), .92 for the Perceptual Reasoning Index (PRI), .92 for the Working Memory Index (WMI), .88 for the Processing Speed Index (PSI), and .97 for the FSIQ. The test-retest coefficients for the WISC-IV are .93 for the VCI, .89 for the PRI, .89 for the WMI, .86 for the PSI, and .93 for the FSIQ. The WISC-IV also

has supported construct validity via exploratory factor analysis as well as confirmatory factor analysis (Flanagan & Kaufman, 2004).

The DAS-II (Elliott, 2007) is an intelligence assessment intended for use with children between the ages of 2 years 6 months and 17 years 11 months. The DAS-II consists of 20 subtests and is broken into two batteries. The School-Age battery is intended for children ages 5:0 to 17:11. The Early Years battery is intended for children aged 2:6 to 6:11 and is further broken into two levels to differentiate between varying skill levels seen in children. The 20 subtests are grouped into Core and Diagnostic subtests.

There are several strengths to the DAS-II. The DAS-II core subtests create the General Conceptual Ability score, which is an estimate of an individual's overall abilities. This score is based on verbal comprehension, fluid reasoning ability, and visual-spatial thinking ability, which are believed to be key mechanisms of thinking ability (Dumont, Willis, & Elliott, 2009). The DAS-II also provides a Special Nonverbal Composite. This composite is generated from performance on nonverbal core subtests. This is helpful when assessing the strengths and weaknesses of nonverbal individuals or those who have expressive language delays. Scoring rules on the DAS-II tend to be clear in the manual and are generally dichotomous (possible score of 0 or 1), with some items being polytomous (scores ranging from 0, 1, 2, or 3), adding to the ease of administration.

A large difference between the DAS-II and many other cognitive assessments is the application of item response theory (Baker, 2001), specifically the Rasch model (Wu & Adams, 2007), which results in the testing being less strenuous for the examinee. In the Rasch model, items are scaled based on how many items were correct, and the relative ability of the individual can be assessed. This is in contrast to other tests, where items may be scaled based on difficulty.

Because it is not necessary to get many items correct in a row to demonstrate a basal, or many items incorrect in a row to demonstrate a ceiling, the test can be less rigorous and tiring for the test taker.

The DAS-II has an average internal consistency reliability coefficient of above .90 for all three batteries. In the area of test-retest reliability, coefficients ranged from .92 in the General Conceptual Ability to .81 for the Nonverbal Reasoning Cluster. The DAS-II also correlates very well with other measures of intelligence ranging from a .88 correlation with the original DAS to a .59 correlation with the Bayley-III Scales of Infant and Toddler Development.

The WJ III COG (Woodcock et al., 2001a) is an intelligence assessment intended for use with individuals from age 2 to the very elderly (norms include individuals in their 90s). It provides a General Intellectual Ability (GIA) score and utilizes a computer program to complete a child's profile. The standard battery contains 10 subtests and the extended battery contains seven additional subtests. The subtests combine to provide the GIA, a Standard or Extended GIA, and Bilingual Intellectual Ability. The overall ability score determined is based on the number and type of subtests analyzed.

The WJ III COG has a standardization sample of over 8,000 and has a strong theoretical foundation (Schrank, Miller, Wendling, & Woodcock, 2010). Basal and ceiling rules are printed in the test manual as well as the test record form, and the tests can be administered in any order. The WJ III COG requires a computer program to transform raw scores to standard scores, which minimizes the error made by users when converting raw scores to standard scores. Median test reliability statistics are provided for the WJ III COG and range from .74 for the Planning subtest to .96 for the Pair Cancellation subtest. Reliability for the different GIA scores ranges from .97 for the GIA–Bilingual, .98 for the GIA–Extended, and .97 for the GIA–Standard. The GIA

scores correlate well with overall scores from other cognitive assessments, ranging from .67 with the WAIS to .76 with the Stanford-Binet Intelligence Scale–Fourth Edition (Schrank et al., 2010).

Definition of Error

Adherence to the instructions provided in the individual manual is essential for obtaining accurate scores on standardized assessments. Failure to adhere to the rules and procedures in the manual can greatly affect raw scores, scaled scores, and index scores, which affects the effectiveness with which a practitioner may utilize the data to properly assist those who are tested. For example, examiner error may significantly change the FSIQ, which will affect IQ-based determinations for the individual.

Errors found in cognitive assessments can be broken into two groups: type and source. The type of error refers to whether the error is due to flawed administration, incorrect scoring, or failure to record information. The source of error refers to whether an error is manualized (a direct noncompliance to a sentence printed in a test manual), arithmetic, or best practices (while not manualized, it is understood that a failure to comply will result in a flawed administration). All errors can be coded for both type and error. Although a large number of errors can be generalized across the different types of cognitive assessment, errors that are specific to a certain type are also examined. General as well as test-specific errors were examined in this study.

Type of Errors

Administration errors are errors related to the specific administration rules or procedures as delineated by the individual test manual. For example, the WISC-IV delineates the appropriate start point for all subtests by age. The Vocabulary subtest indicates that individuals between the ages of 6 and 8 begin at Item 5, individuals between the ages of 9 and 11 begin at

Item 7, and individuals between the ages of 12 and 16 begin at Item 9. Examiners who begin at Item 1, in this case, would be making an administrative error, as they did not begin at the appropriate start point.

Scoring errors are any errors that result in incorrect totals for raw, standard, or index scores. For example, errors included incorrectly tabulating a raw score, incorrect conversion of raw to standard score, and using incorrect scores (e.g., using Sum of Scaled Scores instead of Index Score) when engaging in analysis.

Recording errors are errors that resulted from a failure to record information or verbatim responses. For example, failure to record the completion time of timed items is a recording error that can result in errors calculating the raw score. Failure to record examinee response verbatim was counted as a recording in this study.

Source of Errors

A manualized error is an error that specifically deviates from the instruction provided in a test's manual. All these errors are able to be traced to a specific page within the manual. For example, the WAIS-IV manual instructs individuals to administer sample items on Block Design, Similarities, Digit Span, Matrix Reasoning, Arithmetic, Symbol Search, Visual Puzzles, Coding, Letter-Number Sequencing, Figure Weights, Cancellation, and Picture Completion subtests. Therefore, failure to administer the sample items on these subtests is direct noncompliance with the manual, and is thus considered a manualized error.

An arithmetic error stems from a fundamental error in basic math. For example, failure to add scores to arrive at a raw score is an arithmetic error, as is failure to calculate examinee test age correctly.

A best practices error results from a failure to adhere to administration or recording procedures that will make clinical judgment and scoring choices clear to other individuals examining the protocol. For example, failure to record examinee responses verbatim is a best practices error. Standardized tests such as the WISC-IV are created so that children can be administered the same test in the same fashion no matter where they reside or who tested them. The only way to prove the test's validity is to record verbatim responses to show that the examiner scored the response in the correct way.

Each of the errors described above is, for the purpose of this study, defined as a general errors, because they were found on all the tests examined. Each of the three tests included in the study contained not only these general errors but each was evaluated for specific errors. Specific errors, described below, are errors associated only with the specific test.

Test-specific errors on the DAS-II are errors found only on DAS-II tests and cannot be generalized to other assessments. Examples of test-specific errors on the DAS-II protocols included use of an inappropriate set for the individual's age or ability, failure to utilize a full set, and failure to jump to the next block when the first item is correct.

Test-specific errors on the Wechsler series are errors found only on Wechsler series of tests and cannot be generalized to other assessments. Examples of test-specific errors on the Wechsler series included failure to calculate the longest digit span, subtraction of incomplete items from the total correct item count (e.g., the subtests Cancellation, Symbol Search, and Coding), and inclusion of optional subtests when calculating the sum of scaled scores.

Test-specific errors on the WJ III COG are errors found only on WJ III COG test and cannot be generalized to other assessments. Examples of test-specific errors on the WJ III COG included failure to indicate that corrective feedback had been given, recording the number of

correct items when instructed to record the number of errors, not observing the page rule, and totaling all errors instead of errors per section as instructed.

Appendix A lists general errors, Wechsler-specific errors, WJ III COG-specific errors, and DAS-II-specific errors. Appendix B lists general errors coded by type and source. Appendix C lists all errors per subtest in the three types of tests examined coded by type and source.

For the purpose of this study, subsequent errors that were the sole result of an initial error were not counted. The initial error was counted, but calculation errors that resulted from this initial error were not counted. For example, if the examiner incorrectly recorded a raw score, and thus the standard score was not the correct score as reflected by an individual's performance, the incorrect raw score was counted as an error, but the incorrect standard score was not counted as an error. Additionally, due to the high degree of variability within a given testing session, recalculated FSIQs were not conducted. For example, failure to query may result in incorrect scoring of an item or premature discontinuation of a subtest. This may affect the raw score of a subtest, making it difficult to determine how different the FSIQ should be from what the assessor calculated.

Procedure

The course instructor required master's and doctoral-level students to administer and score four protocols within the Wechsler series, four WJ III COG protocols, and four DAS-II protocols throughout the semester. All protocols were checked for administration, scoring, and recording errors by the teaching assistant and checked again by the course instructor. This descriptive study calculates the frequency and types of examiner errors most commonly committed on the Wechsler series, WJ III COG, and DAS-II.

Scoring checklists that contained general and test-specific errors were developed for the WISC-IV, WAIS-IV, WJ III COG, and DAS-II. The checklists provide information about the number, type, and frequency of administration, recording, and scoring errors committed by student examiners.

Chapter 3

Results

General Error

Examiner errors have been delineated into three areas: administration, computation, and recording. Administration errors are errors related to the specific administration rules or procedures as delineated by the individual test manual. For example, these errors include an inappropriate start point, failure to query the examinee when instructed by the manual, or incorrect use of basal rules. *Computational errors* are any errors that resulted in incorrect totals for raw, standard, or index scores. For example, errors include incorrectly tabulating a raw score, incorrect conversion of raw to standard score, and using incorrect scores (e.g., using Sum of Scaled Scores instead of Index Score) when engaging in analysis. Clerical errors resulting in computation errors were also recorded. These included incorrectly recording dates of birth or dates of administration, and incorrectly calculating age. *Recording errors* are errors that result from a failure to record information or a verbatim response. For example, these errors include failure to record responses verbatim, failure to record completion time or timed items, or failure to keep response booklets or drawings.

General errors made by new and experienced testers were examined. General errors was defined as errors that could have occurred on any of the tests utilized. General errors were first examined by type and source. When errors were grouped by either type or source, a clear pattern emerged. Table 2 shows the number and percentage of errors committed by new and experienced testers, the breakdown by the type, and the source of those errors. With regard to the type of error made, and regardless of the experience level of the examiners, almost half (48% to 59%) of the errors were classified as administrative in nature. For the other types of errors

(Scoring and Recording), there appeared to be an almost even split between them. For example, for the 406 protocols completed by master's level students, approximately 28% were scoring errors and 20% were recording errors. This pattern of an even split between scoring and recording errors for a majority of administrative errors was similar among the three testing groups. When the source of the general errors was examined, manualized errors were most prominent, accounting for between 67% and 78% of all source errors. The remaining source of errors (arithmetic and best practices) was fairly evenly split for the master's and PhD examiners. However, the practitioners appeared to make far more best practices errors (18%) than they did arithmetic errors (4%). For the three groups, the practitioners made more manualized errors (78%) than the other two groups (67% and 69%, respectively), but also made fewer arithmetic errors (4%) than the other two groups (17% and 15%, respectively).

Table 2
Frequency (Number and Percentage) of General Errors by Type and Source From Protocols Completed by New or Experienced Testers

	New testers					
	Master's ($N = 406$)		PhD ($N = 364$)		Practitioners ($N = 40$)	
General error	NE	%	NE	%	NE	%
Type						
Administrative	613	51	557	48	78	59
Scoring	341	28	328	28	23	17
Recording	245	20	285	24	31	24
Source						
Manualized	799	67	810	69	103	78
Arithmetic	205	17	173	15	5	4
Best practices	195	16	187	16	24	18

Note. Percentages are rounded to the nearest whole number. NE = number of errors. *N* refers to the number of protocols available for error review.

Regardless of the type and source of errors, all protocols were examined for the specific errors made. Twenty-seven specific errors had been defined (see Appendix B), and the frequency of those errors in the protocols of all the examiners was tabulated. Table 3 shows the percentage of test records, by all groups combined as well as by the three individual groups, that

had any of the 25 possible general errors. The table also shows the rank order (top 10) of the specific errors by each group. A review of the general errors found that the most frequently occurring errors across all groups was failure to record responses verbatim (43%), incorrect calculation of raw score (38%), and failure to administer sample/practice/teaching items (28%). At least 40% of the errors in all three groups were failure to record response verbatim, making it the most common error for most of the groups. Only in the master's student group did incorrect calculation of raw score happen more frequently (42%) than the error of failure to record responses verbatim. Across the three groups examined, of the 25 possible general errors that could have been made, approximately 12 errors occurred infrequently (less than or equal to 10%).

While the master's and PhD students displayed similar patterns in their commission of errors, for the practitioners a different pattern emerged. Incorrect calculation of raw score, while the second most common error overall, was the ninth most common error for the practitioners. Instead, failure to record responses verbatim (58%, rank order 1), failure to query when instructed by the manual (55%, rank order 2), failure to administer sample/practice/teaching items (43%, rank order 3), inappropriate start points (40%, rank order 4), and failure to query items with specific query criteria (33%, rank order 5) were the most common errors for the practitioners. This observed pattern of practitioners making fewer arithmetic errors and more manualized errors is consistent with prior studies found in the literature.

Table 3

Frequency (%) and Rank Order of Specific General Errors Made by the Total Sample and by New or Experienced Testers

	Total	Master's		PhD		Practitioners	
General errors	%	%	RO	%	RO	%	RO
Failure to record responses verbatim	43	40	2	45	1	58	1
Incorrect calculation of raw score	38	42	1	37	2	13	9
Failure to administer sample/practice/teaching items	28	26	3	28	3	43	3
Failure to stop subtest upon reaching ceiling	23	24	4	25	4	0	
Incorrect analysis of strengths and weaknesses	23	22	6	23	5	23	6
Inappropriate start point	22	23	5	20	6	40	4
Failure to query the examinee when instructed by the manual	18	15	9	19	7	55	2
Failure to establish basal	16	16	7	17	8	5	
Failure to utilize correct format for birth date or administration date	14	12		17	8	20	8
Incorrect instruction or incorrect materials given to the examinee	14	16	7	12		13	9
Failure to reach ceiling	14	14	10	14	10	5	
Incorrect point value assigned to a response	12	7		17		23	6
No query on items with specific query criteria	11	9		11		33	5
Failure to keep response booklets or drawings	7	8		6		3	
Failure to record time of day when indicated	6	3		10		0	
Failure to calculate age correctly	5	5		6		0	
Incorrect conversion of raw score to scaled score/ability score	3	4		3		0	
Incorrect conversion of ability score to T score	3	4		4		0	
Incorrect addition of scaled scores/T scores	2	2		2		0	
Addition of scores for practice items in raw score when not in manual	1	1		2		0	
Incorrect conversion to Index Score/Standard Score	1	1		1		0	
Failure to administer all trials of a multitrial item	0	1		0		0	
Incorrect time limit imposed on the examinee	0	1		0		0	
Failure to record completion time on timed subtests	0	0		0		0	
Failure to reverse/return to previous when basal not establish	0	0		1		0	

Note: All percentages were rounded to the nearest whole number. RO stands for rank order.

Specific Error

Wechsler scales. Master's, PhD, and practitioner protocols from either WISC-IV or WAIS-IV administrations were examined separately from the other two tests (DAS-II and WJ III COG) used in this study. Almost 300 WISC-IV and/or WAIS-IV protocols from master's (N = 141), PhD (N = 114), and practitioners (N = 40) were examined for those tests' specific errors. Because the only protocols received from practitioners were those of the WISC-IV and because 100% of those WISC-IV protocols only contain the 10 core subtests, errors on only those 10 subtests could be examined and tabulated. Each of the 10 core subtests on the WISC-IV is found on the WAIS-IV, although they may not be considered core subtests on the WAIS-IV. Although master's and PhD students did provide WISC-IV as well as WAIS-IV protocols, to be consistent across educational levels, only common subtests across Wechsler scales were compared for this study.

Table 4 presents the percentage of occurrence and cumulative percentages of specific errors made by new or experienced testers found on Wechsler subtests. Given the results found in Table 4, it appears that regardless of the experience of the examiner, errors are the norm. Only one (3%) of the record forms completed by practitioners was totally error-free, whereas for the master's and PhD students, approximately 10% of the record forms evaluated were error-free. The Wechsler protocols administered by master's students (N = 141), PhD students (N = 114), and practicing school psychologists (N = 40) revealed that, regardless of education level, most protocols (between 75% and 86.5%) contained between 0 and 6 errors. Overall, protocols completed by master's students had between 0 and 13 errors, protocols completed by PhD students had between 0 and 22 errors, and protocols completed by practicing psychologists had between 0 and 14 errors. Additionally, 75% of all the protocols completed by master's and PhD-

level students contained between 0 and 4 errors, whereas approximately 75% of protocols completed by practitioners contained between 0 and 6 errors.

Table 4
Frequency (%) and Cumulative Percentage of Specific Errors Made by New or Experienced Testers Found on Wechsler Subtests

	Master's ($N = 141$)		PhD ($N = 114$)		Practitioner ($N = 40$)	
Total errors	%	Cumulative %	%	Cumulative %	%	Cumulative %
0	11	11	10	10	3	3
1	23	34	18	28	15	18
2	21	55	17	45	8	26
3	11	66	18	63	25	51
4	11	77	15	78	10	61
5	7	84	3	81	15	76
6	3	87	4	85	0	76
7	3	90	4	89	5	81
8	4	94	5	94	3	84
9	2	96	0	94	10	94
10	3	99	1	95	0	94
11	1	100	2	97	5	99
12	0	100	0	97	0	99
13	1	100	0	97	0	99
14	-	-	0	97	3	100
15	-	-	0	97	-	-
16	-	-	1	98	-	-
17	-	-	1	99	-	-
18	-	-	1	100	-	-
19	-	-	0	100	-	-
20	-	-	1	100	-	-
21	-	-	0	100	-	-
22	-	-	1	100	-	-

Note: All percentages were rounded to the nearest whole number. Due to this rounding, cumulative percentages did not add up precisely to 100. *N* refers to the number of Wechsler (WISC-IV and WAIS-IV) protocols examined. New testers were the master's and PhD students.

To further examine where errors were made on the Wechsler tests, errors on each individual subtest were tabulated. Table 5 presents the percentage of protocols with at least one error made by master's level, PhD level, and practitioners on specific Wechsler subtests. The subtests that encompass the Perceptual Reasoning Index (PRI) revealed a significant difference between the number of individuals making errors and the educational level of the subjects. The subtests included in the PRI revealed higher percentages of protocols with at least one error

when administered by practitioners. For the 10 core Wechsler subtests examined, only three (Block Design, Picture Concepts, and Matrix Reasoning) showed significant differences in the number of errors made based upon experience level ($x^2 = 7.186, p = .03$; $x^2 = 42.268, p \leq .01$; $x^2 = 18.371, p = .001$, respectively). For these three subtests, the practitioners made significantly more errors than did the less experienced examiners. One additional subtest, Digit Span, had a significant difference between the educational level of the examiner and the number of individuals making errors ($x^2 = 8.290, p = .0158$), with the PhD students making significantly more errors (30%) than the master's group (15%) or the practitioner group (23%).

On Block Design, approximately 50% of the protocols completed by master's and PhD students had at least one error whereas 73% of the practitioner protocols had at least one error. On Picture Concepts, approximately 3% to 4% of the protocols completed by master's and PhD students had at least one error, whereas 33% of the practitioner protocols had at least one error. Finally, on Matrix Reasoning, approximately 17% to 25% of the record forms completed by master's and PhD students had at least one error whereas 50% of the practitioner record forms had at least one error. This result indicated that practitioners, as a group, made more errors on these subtests than did the inexperienced students.

In contrast, on the Digit Span subtest, approximately 15% and 23% of the protocols completed by master's students or practitioners, respectively, had at least one error, whereas almost one in three (30%) of the record forms completed by PhD students had one or more errors.

Table 5
Percentage of Protocols With At Least One Error Made by New or Experienced Testers on Specific Wechsler Subtests

	% of protocols with at least 1 error		
Domain/subtest	Master's	PhD	Practitioner
Verbal Comprehension			
Similarities	40	43	60
Vocabulary	38	46	58
Comprehension	42	49	33
Perceptual Reasoning			
Block Design	50	49	73
Picture Concepts	3	4	33
Matrix Reasoning	17	25	50
Working Memory			
Digit Span	15	30	23
Letter-Number Sequencing	16	25	20
Processing Speed			
Symbol Search	20	11	15
Coding	2	2	5

Note. All percentages were rounded to the nearest whole number. New testers were the master's and PhD students.

The following analysis relating to the DAS-II and the WJ III COG only included protocols provided by master's and PhD students. None of the protocols obtained from practitioners were DAS-II or WJ III COG. Because of the lack of practitioners as a comparison group, analysis of errors on DAS-II and WJ III COG focused only on the combined new testers (master's/PhD) groups. Because no comparison was made to the practitioners, error analysis on the DAS-II and the WJ III COG included not only specific and general errors, similar to the Wechsler scales, but also analysis of the number of possible subtest errors compared to actual subtest errors.

DAS-II. Errors on 257 DAS-II protocols were examined, and tabulations were made of all errors. Table 6 presents the percentage of occurrence and cumulative percentages of general and specific errors found on the DAS-II made by new testers. The protocols examined had between zero and 15 errors, and approximately 75% of the protocols had between zero and five errors.

Table 6
Frequency (%) and Cumulative Percentage of General and Specific Errors Found on the DAS-II Made by New Testers

Total errors	%	Cumulative %
0	11	11
1	16	27
2	19	46
3	14	60
4	7	67
5	10	77
6	10	87
7	4	91
8	3	94
9	4	98
10	2	100
11	<1	100
14	<1	100
15	<1	100

Note. All percentages were rounded to the nearest whole number.

To examine further where errors were made on the DAS-II, errors on each individual subtest were tabulated. Table 7 delineates, for each DAS-II subtest, the possible number of errors that could be made, the range of errors that were actually committed, and the percentage of protocols with errors. Table 7 also includes the percentage of errors that each subtest accounted for in the total number of errors possible and the percentage of errors on each subtest that actually occurred. For example, the Word Definition subtest had a total of nine possible specific errors that could be made by any examiner, but on the 257 protocols reviewed, only zero to three errors were made. Twenty-three percent of all the protocols reviewed had errors on the Word Definition subtest. The number of possible errors on the Word Definition subtest accounted for 8% of the total possible errors on the DAS-II. The errors found on the Word Definition subtest on all the protocols examined accounted for 8% of the total errors, which was perfectly in line with the expectation.

Table 7
Number of Possible Errors, Range of Errors, Percentage of Protocols With Zero Errors, and the Percentage of Errors Possible/Occurred on the DAS-II

Subtest	Subtest errors		% of protocols	%	
	Possible	Range	with errors	Possible	Occurred
Core					
Word Definitions	9	0-3	23	8	8
Verbal Similarities	9	0-2	24	8	7
Matrices	10	0-4	24	9	8
Sequential and Quantitative Reasoning	9	0-1	15	8	4
Recall of Designs	10	0-3	35	9	13
Pattern Construction	10	0-3	50	9	15
Diagnostic					
Recall of Objects—Immediate	3	0-1	1	3	0
Recall of Objects—Delayed	3	-	0	3	0
Recall of Digits Forward	9	0-2	32	8	10
Recognition of Pictures	10	0-2	22	9	7
Recall of Sequential Order	13	0-3	47	11	17
Recall of Digits Backwards	12	0-2	21	10	10
Phonological Processing	5	0-1	1	4	0
Rapid Naming	5	0-3	3	4	1

Note. All percentages were rounded to the nearest whole number.

Of the 14 subtests included in the tabulation of errors, the majority of subtests had no errors at all. For example, for Recall of Objects—Delayed, although there were three possible errors that could have been made, none of the protocols examined had any one of those three errors. For the subtest Recall of Digits Backwards, where there was a possibility of 12 errors, 69% of the protocols reviewed had no errors at all. However, results did reveal two problematic subtests. About 50% of the protocols examined showed errors for the Pattern Construction and Recall of Sequential Order subtests. Although Pattern Construction and Recall of Sequential Order should have accounted for 9% and 11% of the total errors, respectively, they actually accounted for 15% and 17%, respectively.

WJ III COG. Errors on the 258 WJ III COG protocols were examined, and tabulations were made of all errors. Table 8 presents the percentage of occurrence and cumulative percentage of general and specific errors found on the WJ III COG made by new testers. The

protocols examined had between 0 and 33 errors, and 75% of the protocols had between 0 and 18 errors.

Table 8
Frequency Distribution, Percentage, and Cumulative Percentage of Total Errors Made on the WJ III COG

Total errors	Count	%	Cumulative %
0	23	9	9
1	21	8	17
2	24	9	26
3	19	7	33
4	12	5	38
5	10	4	42
6	13	5	47
7	12	5	52
8	6	2	54
9	6	2	56
10	8	3	59
11	1	0	59
12	5	2	61
13	3	1	62
14	4	2	64
15	2	1	65
16	8	3	68
17	8	3	71
18	11	4	75
19	13	5	80
20	11	4	84
21	13	5	89
22	3	1	90
23	2	1	91
24	3	1	92
25	5	2	94
26	5	2	96
27	1	0	96
28	1	0	96
29	2	1	97
30	1	0	97
32	1	0	97
33	1	0	97

Note. All percentages were rounded to the nearest whole number. Due to this rounding, the cumulative percentage did not add up precisely to 100.

To examine further where errors were made on the WJ III COG, errors on each individual subtest were tabulated. Table 9 delineates, for each WJ III COG subtest, the possible number of errors that could be made, the range of errors actually committed, and the percentage of

protocols with errors. Table 9 also includes the percentage of errors that each subtest accounted for in the total number of errors possible, and the percentage of errors on each subtest that actually occurred. For example, the Verbal Comprehension subtest had a total of nine possible specific errors that could be made by any examiner, but on the 258 protocols reviewed, only zero to three errors were made. Fifty-eight percent of all the protocols reviewed had errors at all on the Verbal Comprehension subtest. The number of possible errors on the Verbal Comprehension subtest accounted for 8% of all the total possible errors on the WJ III COG. In this case, the errors found on the Verbal Comprehension subtest on all the protocols examined accounted for 8% of the total errors, which is perfectly in line with the expectation.

Table 9
Number of Possible Errors, Range of Errors, Percentage of Protocols With Zero Errors, and the Percentage of Errors Possible/Occurred on the WJ III COG

	Subtest errors			%	
Subtest	Number possible	Range	% of protocols with errors	Possible	Occurred
Verbal Comprehension	9	0-3	58	8	8
Visual-Auditory Learning	7	0-5	65	7	13
Spatial Relations	5	0-2	41	5	4
Sound Blending	5	0-2	42	5	5
Concept Formation	5	0-2	41	5	4
Visual Matching	4	0-1	3	4	0
Numbers Reversed	7	0-4	52	7	7
Incomplete Words	5	0-2	51	5	6
Auditory Working Memory	6	0-3	50	6	6
Visual-Auditory Learning Delayed	7	0-4	60	7	11
General Information	8	0-3	55	7	7
Retrieval Fluency	2	0-0	0	2	0
Picture Recognition	5	0-2	42	5	5
Auditory Attention	5	0-2	44	5	5
Analysis-Synthesis	5	0-2	44	5	5
Decision Speed	3	0-1	1	3	0
Memory for Words	7	0-3	43	7	5
Rapid Picture Naming	4	0-2	32	4	3
Planning	5	0-3	48	5	7
Pair Cancellation	3	0-2	0	3	0

Note: All percentages were rounded to the nearest whole number.

For 258 WJ III COG protocols examined, errors for each of the 20 subtests were independently tabulated. On two of the 20 subtests (Retrieval Fluency and Pair Cancellation) on

which examiners had the possibility of making two or three errors, respectively, no errors were found. However, results did reveal some problematic subtests. For example, the subtests with the lowest percentage of error-free performance were Visual Auditory Learning and Visual Auditory Learning—Delayed, where 65% and 60% of the protocols examined contained errors. Although both Visual Auditory Learning and Visual Auditory Learning—Delayed should each have accounted for only 7% of the total errors, they actually accounted for 13% and 11%, respectively.

Chapter 4

Discussion

Although errors on various cognitive tests have been examined before, no studies were found that compared the number and type of errors made by new versus experienced testers, and no studies were found that compared the number and kind of errors made by the same sample on different cognitive tests. Previous studies typically concentrated on only one aspect of testing error, for example on errors made by graduate students only or errors made by practicing psychologists only. Previous studies also examined only one test, and no studies included the DAS.

Upon examining the results of this study, it is clear that errors on cognitive assessments are the norm, not the exception, regardless of educational level or type of assessment. It is a concerning trend that errors persist regardless of training or even after finishing an educational program. In this study, errors were divided into type and source and an examination ensued regarding which categories these errors fell into. However, a more simplistic view of the error delineation could be termed as real versus best practice error. For example, an incorrect calculation of a raw score is a real error, as it is very clearly spelled out and delineated in the test manual. In contrast, failure to record responses verbatim is a best practice error because failure to record responses does not necessarily have a significant impact on the final scores, although best practice would suggest that it be done. In this study, it was believed that best practice was a valuable aspect of error making to be examined. The results of this study revealed that best practice errors are, in fact, important for all test administrators to be mindful of because this type of error was commonly found across all tests. Professionals should all strive for best practice, as it ensures defendable results. Although best practice errors are not as clearly spelled out in the

test manuals, they do in fact have the potential for greatly reducing the validity of test results. As noted earlier, the practical implications of this type of error could be an inability to justify how one obtained a certain score on an item or an entire subtest, as well as difficulty in defending one's results in a legal situation.

Practitioner Error

The results obtained from this study are consistent with the results reported previously and indicate that errors made on standardized cognitive assessments by graduate student as well as practitioners are the norm rather than the exception. Consistent with prior research, failure to record responses verbatim was the most common error (Alfonso et al., 1998; Belk et al., 2002; Loe et al., 2007). Surprisingly, the current study also revealed that, overall, experienced testers often made more errors than the student evaluators. This finding, in and of itself, has specific implications for the field of standardized assessment. Practitioners need to be reminded not to rely only on their memory for the scoring of verbal responses and to double check their administration procedure, tabulation of raw scores, and so forth, in order to maintain accurate scoring. Practitioners administering the WISC-IV often made errors relating to not recording responses verbatim, utilizing incorrect start points, failing to administer practice items, and failing to query the examinee. The practitioners did not make as many arithmetic errors as graduate students, indicating that they were well-versed in how to administer the measure, but made errors in favor of minimizing time spent giving the test. For example, practitioners were often found to not be utilizing correct start points, instead beginning the test with more difficult items. This may have happened because it was inferred that the examinee knew the initial items and the practitioner did not want to give the time for these obvious answers. Although this reasoning may seem logical, it goes against the administrative directions in the manual. Another

common error among practitioners was to give point values to responses without querying where the manual makes clear a query is needed. This is possibly due to examiners not actively looking at a manual during administration and the examiners' belief that they were correctly awarding the points or relying on memory. Overall, this points to a failure to complete record forms with accuracy and best practices, and this finding is consistent with Erdodi et al. (2009), who suggested that practitioners may utilize their memory rather than using the manual to score responses. Continued training after certification may assist in ensuring that these bad habits be minimized in favor of error-free administration.

Errors on All Assessments

The most common errors found for individuals learning to administer cognitive assessment as well as practicing school psychologists were that they used incorrect start points and did not establish correct basals and ceilings according to the rules set forth in the manual. It often appeared as if individuals attempted to generalize an overall start point/basal/ceiling rule, such as beginning with the first item, or the first three correct items being the basal, or the last three incorrect items being the ceiling. Individuals administering cognitive assessments are encouraged to rely on the manual as well as the test record, as both of these sources provide start point, basal, and ceiling information for every subtest. Rectifying these errors for individuals administering cognitive assessments can go a long way toward reducing manualized error.

The trend of the majority of general errors being administrative or manualized may be partially explained by possible administrative errors being more numerous than other types of errors and manualized errors being the more numerous source of errors. Therefore, it was more likely for any given error to be administrative or manualized than any other type.

Additionally, incorrect computation of the raw score was a very common error. Although this was often a mathematical error, there also seemed to be confusion regarding whether or not to include unadministered items in the overall raw score. Individuals learning to administer cognitive assessments should look to the test manual for correct methods for obtaining the raw scores.

Wechsler

The Wechsler series contains subtests that are often very different from each other, frequently with differing basal and ceiling rules. This results in errors when individuals are learning the specifics of each subtest. Additionally, lengthy basals and ceilings may make the administration of a subtest very long and frustrating for the test taker, as when the examinee continues to get one item correct in a long string of incorrect answers, only to be given more questions as the ceiling rule was not met.

A positive aspect of the Wechsler series is the designated areas to record examinee responses. This allows examiners to more easily complete record forms with best practices.

DAS-II

The format of the DAS-II is significantly different from the other two cognitive assessments examined, mainly in the use of sets. Additionally, subtests where a less than maximum score can dictate whether to continue to the next decision point, or where less than three first-trial passes can dictate whether to return to a previous start point, can be confusing and appear to require demonstration from a knowledgeable practitioner to master. In the DAS-II, basals and ceilings are also not necessarily in the format of three maximum scores in a row or three failures in a row, as found in other cognitive assessments. This also appears to require additional clarification for individuals learning to administer cognitive assessments.

Additionally, the DAS-II utilizes sets rather than ceilings in the format of three incorrect in a row. Individuals appear to have initial difficulty utilizing a set as delineated by the manual, and this may require support by the class instructor.

The DAS-II also is not consistent with whether or not unadministered items are to be included in the raw score. On some subtests, only items included in the set are counted, as in Recall of Designs, whereas on other subtests with block jumping, unadministered items are counted toward the raw score, as in Recall of Sequential Order. The differing rules of the subtests in the DAS-II may have accounted for a large number of the errors. Overall, the use of sets and blocks in the DAS-II allows examinees to reach a ceiling without the need for several wrong answers in a row, which not only reduces test time but also negatively affects the test taker's mood.

WJ III COG

The WJ III COG record forms have fairly consistent basal and ceiling rules, making learning the test easier when compared to other assessments. However, a significant general trend on the WJ III COG was failure to record responses verbatim. The record form provides limited designated room to record responses, which affects the administrative performance of individuals completing record forms with best practices. The WJ III COG had the highest number of errors in a single protocol out of all protocols examined, and often when an examiner failed to record a response, it was done universally throughout the record form. This increases the number of errors found, as failure to record a response in each subtest would count as one error per subtest.

However, the WJ III COG has fairly consistent basal and ceiling rules throughout the assessment, as well as consistent methods to obtain a raw score. This makes arithmetic errors less likely and allows the examiner to easily obtain basals and ceilings.

Recommendations for Assessment Courses

For the purposes of the graduate-level cognitive assessment course, students were not required to wait for feedback before handing in another record form to be graded. As such, all four record forms of one assessment were often handed in at the same time, resulting in occurrences of the same error with no opportunity for feedback and no demonstration of improvement. This is concerning, because students often make different errors after feedback and those that do not have an opportunity for additional feedback may not remedy these errors as a practitioner. Therefore, a recommendation is to require students to wait for feedback before handing in another record form to be graded. In this way, evaluators who make mistakes without knowing they are mistakes do not repeatedly make the same mistakes.

Research has also shown that requiring students to maintain a portfolio of protocols to review before administration significantly reduced error (Egan et al., 2003). This recommendation may be helpful to students, as it will allow them to recheck their previous errors before administering another assessment.

Recommendations for Practitioners

Practitioners are always encouraged to strive for best practices, and the results of this study indicated that errors are very common among practitioners. In fact, the results of this study revealed that only 1% of practitioner-completed protocols were error-free, compared to 10% of the student-completed protocols. Given the impact that incorrectly completed protocols can have on the overall performance profile of an individual, this is highly problematic when looking at

the different areas of current U.S. policy that utilize cognitive assessment. Practitioners are urged to continue education in current assessments and to continuously strive for error-free completion of protocols. Practitioners are also urged not to rely on memory to score responses and to always check whether a response warrants query.

Recommendations for Future Research

This study did not examine the impact of examiner error on derived standard scores. Continued study in this area is a recommendation for future research, especially with the relatively unstudied DAS-II. As assessments are updated to current versions, new manuals are made and new test procedures are put in place. Continued research on cognitive assessment errors is recommended, as errors are clearly the norm rather than the exception as demonstrated by this study and all previous studies.

Additionally, this was the first study found in lit review that compared students to practitioners, and the first study that compared different tests. Given that the results of this study revealed significant differences, continued study in cross-educational level and cross-assessment areas would be valuable.

References

Aaron, P., Joshi, R., Gooden, R., & Bentum, K. (2008) Diagnosis and treatment of reading disabilities based on the component model of reading: An alternative to the discrepancy model of LD. *Journal of Learning Disabilities, 41*, 67-84.

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

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

American Psychological Association. (2010). *American Psychological Association ethical principles of psychologists and code of conduct.* Retrieved November 11, 2010, from http://www.apa.org/ethics/code/index.aspx

Atkins v. Virginia, 536 U.S. 304 (2002).

Baker, F. (2001). *The basics of item response theory.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Barnes, M., Fletcher, J., & Fuchs, L. (2007). *Learning disabilities: From identification to intervention.* New York, NY: The Guilford Press.

Belk, M., LoBello, S., Ray, G., & Zachar, P. (2002). WISC-III administration, clerical, and scoring errors made by student examiners. *Journal of Psychoeducational Assessment, 20*, 290-300.

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

Brazelton, E., Jackson, R., Buckhalt, J., Shapiro, S., & Byrd, D. (2003). Scoring errors on the WISC-III: A study across levels of education, degree fields, and current professional positions. *Professional Educator, 25*(2), 1-8.

Brown, M., Swigart, M., Bolen, L., Hall, C., & Webster, R. (1998). Doctoral and nondoctoral practicing school psychologists: Are there differences? *Psychology in the Schools, 35,* 347-354.

Clark, B. (2002). *Growing up gifted.* Columbus, OH: Merrill Prentice Hall.

Cody, M., & Prieto, L. (2000). The teaching of individual mental assessment: A national survey. *The Teaching of Psychology, 27*(3), 190-194.

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

DeMatteo, D., Marczyk, G., & Pich, M. (2007). A national survey of state legislation defining mental retardation: implications for policy and practice after Atkins. *Behavioral Sciences & the Law, 25,* 781-802.

Dumont, R., Willis, J. O., & Elliott, C. D. (2009). *Essentials of DAS-II assessment.* Hoboken, NJ: Wiley.

Egan, P., McCabe, P., Semenchuk, D., & Butler, J. (2003). Using portfolios to teach test scoring skills: A preliminary investigation. *Teaching of Psychology, 30*(3), 233-235.

Elliott, C. D. (2007). *Differential Ability Scales, Second Edition (DAS-II).* San Antonio, TX: Pearson Education.

Ellis, A., Abrams, M., & Abrams, L. (2009) *Personality theories: Critical perspectives.* Thousand Oaks, CA: Sage.

Erdodi, L., Richard, D., & 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.

Flanagan, D., & Kaufman, A. (2004). *Essentials of WISC-IV Assessment.* Hoboken, NJ: Wiley.

Fletcher J. (1992). The validity of distinguishing children with language and learning disabilities according to discrepancies with IQ. *Journal of Learning Disabilities, 25*, 546-548.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Gross, M. (2004). *Exceptionally gifted children* (2nd ed.). New York, NY: RoutledgeFalmer.

Implementation of a Sentence of Death § 18 U.S.C. § 3596 (2007).

Johnsen, S. (2004). *Identifying gifted students: A practical guide.* Waco, TX: Prufrock Press.

Kaplan, N. (2007, August 30). *Want a job at Google? Try these brainteasers first.* Retrieved March 31, 2011, from http://money.cnn.com/2007/08/29/technology/brain_teasers.biz2/index.htm

Karlgaard, R. (2005, October 31). *Talent wars.* Retrieved August 6, 2006, from http://www.forbes.com/forbes/2005/1031/045.html

Kuzmits, F., & Adams, A. (2008). The NFL combine: Does it predict performance in the National Football League? *Journal of Strength and Conditioning Research, 22*, 1721-1727.

Lichtenberger, E. O., & Kaufman, A. S. (2004). *Essentials of WPPSI-III Assessment.* Hoboken, NJ: Wiley.

Lichtenberger, E. O., & Kaufman, A. S. (2009). *Essentials of WAIS-IV Assessment.* Hoboken, NJ: Wiley.

Loe, S., Kadlubek, R., & Marks, W. (2007). Administration and scoring errors on the WISC-IV among graduate student examiners. *Journal of Psychoeducational Assessment, 25*, 237-247.

Lombardo, P. (2011). *A century of eugenics in America.* Bloomington: Indiana University Press.

Miklos, D., & Carlson, E. (2000). Engineering American society: The lesson of eugenics. *Macmillan Magazines, 1,* 153-158.

Murphy, K. (2002). Can conflicting perspectives on the role of g in personnel selection be resolved? *Human Performance, 15*, 173-186.

New Jersey State Board of Education. (2005). N.J.A.C. 6A: 8, Standards and Assessment for Student Achievement.

Oakland, T. D., & Zimmerman, S. A. (1986). The course on individual mental assessment: A national survey of course instructors. *Professional School Psychology, 1*, 51-59.

Orme, D., Brehm, W., & Ree, M. (2001). Armed forces qualification test as a measure of premorbid intelligence. *Military Psychology, 13*(4), 187-197.

Patterson, M., Slate, J., Jones, C., & Steger, H. (1995). The effects of practice administrations in learning to administer and score the WAIS-R: A partial replication. *Educational and Psychological Measurement, 55*, 32-37.

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

Rodger, E. (2011, January 20). A question regarding cognitive assessment training [Message posted to SPTRAIN electronic mailing list]. Retrieved from http://lsv.uky.edu/scripts/wa.exe?A1=ind1101&L=sptrain

Schrank, F. A., Miller, D. C., Wendling, B. W., & Woodcock, R. W. (2010). *Essentials of WJ III Cognitive Abilities Assessment.* Hoboken, NJ: Wiley.

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

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools, 26*, 74-84.

Slate, J. R., & Jones, C. H. (1990). Student error in administering the WISC-R: Identifying problem areas. *Measurement & Evaluation in Counseling & Development, 23*, 137-140.

Slate, J. R., & Jones, C. H. (1993). Evidence that practitioners err in administering and scoring the WAIS-R. *Measurement & Evaluation in Counseling & Development, 20*(4), 156-162.

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

Slate, J. R., Jones, C. H., Murray, R. A., & Coulter, C. (1993). Evidence that practitioners err in administering and scoring the WAIS-R. *Measurement and Evaluation in Counseling and Development, 25*, 156-161.

Social Security Administration. (2008). Disability evaluation under Social Security: 12.00 Mental Disorders – Adult. Retrieved June 6, 2010, from http://www.ssa.gov/disability/professionals/bluebook/.

Stern, A. M. (2007). We cannot make a silk purse out of a sow's ear: Eugenics in the Hoosier Heartland. *Indiana Magazine of History 103*, 3-38.

Sternberg, R. (Ed.). (2004). *Definitions and conceptions of giftedness.* Thousand Oaks, CA: Corwin Press.

Super, J. (2006). A survey of pre-employment psychological evaluation tests and procedures. *Journal of Police and Criminal Psychology, 21*(2), 83-87.

U.S. Department of Defense. (2005, September 20). *Department of Defense instruction, Number 1145.01*. Retrieved from http://www.dtic.mil/whs/directives/corres/pdf/114501p.pdf

Warren, S. A., & Brown, W. G. (1972). Examiner scoring errors on individual intelligence tests. *Psychology in the Schools, 9*, 118-122.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence—Third Edition (WPPSI-III)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children* (4th ed.) (WISC-IV). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003b). *WISC-IV technical and interpretive manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV)*. San Antonio, TX: The Psychological Corporation.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001a). *Examiner's manual. Woodcock-Johnson III Tests of Cognitive Ability*. Itasca, IL: Riverside.

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

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne, Australia: Educational Measurement Solutions.

Appendix A: List of Error Types: General, Wechsler Specific, WJ III Specific, and DAS-II Specific

General Errors

Recording Errors

Failure to utilize correct format for birth date or administration date

Failure to calculate age correctly

Failure to record responses verbatim

Failure to record completion time on timed subtests

Failure to keep response booklets or drawings

Administration Errors

Inappropriate start point

Failure to query the examinee when instructed by manual

Failure to establish basal

Failure to reverse/return to previous point when basal is not established

Failure to reach ceiling

Failure to stop subtest upon reaching ceiling

Failure to administer sample/practice/teaching items

Failure to administer all trials of a multitrial item

Incorrect materials given to the examinee

Incorrect time limit imposed on the examinee

Scoring Errors

Assigning an incorrect point value to a response

Incorrect calculation of raw score

Addition of scores for practice items in raw score when not indicated by manual

Wechsler Specific Errors

Recording Errors

Failure to record time of day when indicated

Administration Errors

Scoring Errors

Incorrect calculation of longest digit span

Inclusion of uncompleted items when calculating incorrect items

Incorrect conversion of raw score to Scaled Score

Incorrect addition of Scaled Scores/T Scores

Incorrect conversion to Index Score/Standard Score

Incorrect analysis of strengths and weaknesses

WJ III Specific Errors

Recording Errors

Failure to indicate when corrective feedback is given

Failure to record time of day when indicated

Administration Errors

Failure to administer full page when indicated by manual

Scoring Errors

Calculation of correct responses instead of incorrect responses on the subtest

Planning

DAS-II Specific Errors

Recording Errors

Failure to record time of day when indicated

Administration Errors

Failure to utilize an existing set

Failure to proceed to first item in next block when first item is correct

Failure to administer whole block if first item in last block is correct

Scoring Errors

Addition of unadministered items to raw score when not indicated by manual

Incorrect conversion of raw score to Ability Score

Incorrect conversion of Ability Score to T Score

Incorrect addition of Scaled Scores/T Scores

Incorrect conversion to Index Score/Standard Score

Incorrect analysis of strengths and weaknesses

Appendix B: General Errors Coded by Type and Source

Types: Administration (A), Scoring (S), Response (R)

Source: Manualized (M), Arithmetic (A), Best Practices (BP)

Error description	Type	Source
1. Failure to utilize correct format for birth date or administration date	R	M
2. Failure to calculate age correctly	R	A
3. Incorrect conversion of raw score to scaled score/ability score	S	M
4. Incorrect conversion of ability score to T score	S	M
5. Incorrect addition of scaled scores/T scores	S	A
6. Incorrect conversion to Index Score/Standard Score	S	M
7. Incorrect analysis of strengths and weaknesses	S	M
8. Failure to keep response booklets or drawings	A	BP
9. Failure to record time of day when indicated	R	M
10. Incorrect calculation of raw score	S	A
11. Failure to administer sample/practice/teaching items	A	M
12. Addition of scores for practice items in raw score when not in manual	S	A
13. Addition of unadministered items to raw score when not indicated	S	M
14. Inclusion of uncompleted items when calculating incorrect items	S	M
15. Inappropriate start point	A	M
16. Failure to record responses verbatim	R	BP

17. Failure to record completion time on timed subtests	R	M
18. Incorrect point value assigned to a response	S	M
19. Failure to establish basal	A	M
20. Failure to reverse/return to previous when basal not establish	A	M
21. Failure to reach ceiling	A	M
22. Failure to stop subtest upon reaching ceiling	A	M
23. Failure to administer all trials of a multi-trial item	A	M
24. Failure to query the examinee when instructed by the manual	A	M
25. No query on items with specific query criteria	A	M
26. Incorrect instruction or incorrect materials given to the examinee	A	M
27. Incorrect time limit imposed on the examinee	A	M

Appendix C: Subtest-Specific Errors in All Tests Coded by Type and Source

Types: Administration (A), Scoring (S), Response (R)

Source: Manualized (M), Arithmetic (A), Best Practices (BP)

WAIS IV/WISC IV

Subtests are in order found on the WAIS-IV, deviations found in the WISC-IV are noted

Error Description	Type	Source
1) Block Design		
1. Incorrect calculation of raw score	S	A
2. Incorrect calculation of raw score with no time bonus	S	A
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A
5. Failure to record responses verbatim	R	BP
6. Assigning an incorrect point value to a response	S	M
7. Inappropriate start point	A	M
8. Failure to establish basal	A	M
9. Failure to reach ceiling	A	M
10. Failure to stop subtest upon reaching ceiling	A	M
2) Similarities		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M

6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
10. Failure to query the examinee when instructed by manual	A	M
3) Digit Span		
1. Incorrect calculation of raw score	S	A
2. Incorrect recording of longest digit span	R	M
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A
5. Failure to record responses verbatim	R	BP
6. Assigning an incorrect point value to a response	S	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
4) Matrix Reasoning (subtest 8 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M

8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
5) Vocabulary (subtest 6 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to record responses verbatim	R	BP
3. Assigning an incorrect point value to a response	S	M
4. Inappropriate start point	A	M
5. Failure to establish basal	A	M
6. Failure to reach ceiling	A	M
7. Failure to stop subtest upon reaching ceiling	A	M
8. Failure to query the examinee when instructed by manual	A	M
6) Arithmetic (subtest 14 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
7) Symbol Search (subtest 10 on WISC)		
1. Failure to record completion time on timed subtests	R	M

2. Examinee not completing items does not count as incorrect	S	M
3. Incorrect calculation of raw score	S	A
4. Incorrect materials/instruction given to the examinee	A	M
8) Visual Puzzles		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
9) Information (subtest 13 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to record responses verbatim	R	BP
3. Assigning an incorrect point value to a response	S	M
4. Inappropriate start point	A	M
5. Failure to establish basal	A	M
6. Failure to reach ceiling	A	M
7. Failure to stop subtest upon reaching ceiling	A	M
8. Failure to query the examinee when instructed by manual	A	M
9. Failure to query on items with specific query criteria	A	M

10) Coding (subtest 5 on WISC)		
1. Failure to record completion time on timed subtests	R	M
2. Examinee not completing items does not count as incorrect	S	M
3. Incorrect calculation of raw score	S	A
11) Letter-Number Sequencing (subtest 7 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Incorrect recording of longest digit span	R	M
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A
5. Failure to record responses verbatim	R	BP
6. Assigning an incorrect point value to a response	S	M
7. Inappropriate start point	A	M
8. Failure to establish basal	A	M
9. Failure to reach ceiling	A	M
10. Failure to stop subtest upon reaching ceiling	A	M
12) Figure Weights		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M

8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
13) Comprehension (subtest 9 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
10. Failure to query the examinee when instructed by manual	A	M
11. Failure to query on items with specific query criteria	A	M
14) Cancellation (subtest 12 on WISC)		
1. Failure to record completion time on timed subtests	R	M
2. Examinee not completing items does not count as incorrect	S	M
3. Incorrect calculation of raw score	S	A
4. Incorrect materials/instruction given to the examinee	A	M
15) Picture Completion (subtest 11 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A

4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
16) Picture Concepts (subtest 4 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
17) Word Reasoning (subtest 15 on WISC)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M

7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M

DAS-2 School Age/Early Years

Subtests are in order found on the School Age protocol, deviations found in the Early Years protocol are noted.

Error description	Type	Source
1) Recall of Designs		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
10. Failure to keep drawings	A	BP
2) Word Definitions		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to record responses verbatim	R	BP

4. Assigning an incorrect point value to a response	S	M
5. Inappropriate start point	A	M
6. Failure to establish basal	A	M
7. Failure to reach ceiling	A	M
8. Failure to stop subtest upon reaching ceiling	A	M
9. Failure to query the examinee when instructed by manual	A	M
3) Recall of Objects – Immediate (subtest 4 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to record time of day	R	M
3. Failure to record responses verbatim	R	BP
4) Pattern Construction (subtest 5 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A
5. Failure to record responses verbatim	R	BP
6. Assigning an incorrect point value to a response	S	M
7. Inappropriate start point	A	M
8. Failure to establish basal	A	M
9. Failure to reach ceiling	A	M
10. Failure to stop subtest upon reaching ceiling	A	M
5) Matrices (subtest 6 on Early Years)		
1. Incorrect calculation of raw score	S	A

2. Failure to utilize an existing set	A	M
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A
5. Failure to record responses verbatim	R	BP
6. Assigning an incorrect point value to a response	S	M
7. Inappropriate start point	A	M
8. Failure to establish basal	A	M
9. Failure to reach ceiling	A	M
10. Failure to stop subtest upon reaching ceiling	A	M
6) Recall of Objects – Delayed (subtest 7 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to record time of day	R	M
3. Failure to record responses verbatim	R	BP
7) Verbal Similarities		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to record responses verbatim	R	BP
4. Assigning an incorrect point value to a response	S	M
5. Inappropriate start point	A	M
6. Failure to establish basal	A	M
7. Failure to reach ceiling	A	M
8. Failure to stop subtest upon reaching ceiling	A	M
9. Failure to query the examinee when instructed by manual	A	M

8) Sequential and Quantitative Reasoning		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to record responses verbatim	R	BP
4. Assigning an incorrect point value to a response	S	M
5. Inappropriate start point	A	M
6. Failure to establish basal	A	M
7. Failure to reach ceiling	A	M
8. Failure to stop subtest upon reaching ceiling	A	M
9) Recall of Digits Forward		
1. Incorrect calculation of raw score	S	A
2. Failure to record responses verbatim	R	BP
3. Assigning an incorrect point value to a response	S	M
4. Failure to establish basal	A	M
5. Failure to reach ceiling	A	M
6. Failure to stop subtest upon reaching ceiling	A	M
7. Failure to jump to first item in next block when first item is correct	A	M
8. Failure to administer whole block if first item in last block is correct	A	M
10) Recognition of Pictures		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A

5. Failure to record responses verbatim	R	BP
6. Assigning an incorrect point value to a response	S	M
7. Inappropriate start point	A	M
8. Failure to establish basal	A	M
9. Failure to reach ceiling	A	M
10. Failure to stop subtest upon reaching ceiling	A	M
11) Recall of Sequential Order (subtest 12 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
10. Failure to jump to first item in next block when first item is correct	A	M
11. Failure to administer whole block if first item in last block is correct	A	M
12. Failure to administer all items in order after Sample F	A	M
12) Recall of Digits Backward (subtest 13 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A

4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
6. Inappropriate start point	A	M
7. Failure to establish basal	A	M
8. Failure to reach ceiling	A	M
9. Failure to stop subtest upon reaching ceiling	A	M
10. Failure to jump to first item in next block when first item is correct	A	M
11. Failure to administer whole block if first item in last block is correct	A	M
13) Phonological Processing (subtest 14 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
14) Rapid Naming (subtest 14 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to administer sample/practice/teaching items	A	M
3. Addition of scores for practice items in raw score	S	A
4. Failure to record responses verbatim	R	BP
5. Assigning an incorrect point value to a response	S	M
15) Verbal Comprehension (subtest 1 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M

3. Failure to record responses verbatim	R	BP
4. Assigning an incorrect point value to a response	S	M
5. Inappropriate start point	A	M
6. Failure to establish basal	A	M
7. Failure to reach ceiling	A	M
8. Failure to stop subtest upon reaching ceiling	A	M
9. Failure to query the examinee when instructed by manual	A	M
16) Picture Similarities (subtest 2 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to record responses verbatim	R	BP
4. Assigning an incorrect point value to a response	S	M
5. Inappropriate start point	A	M
6. Failure to establish basal	A	M
7. Failure to reach ceiling	A	M
8. Failure to stop subtest upon reaching ceiling	A	M
17) Naming Vocabulary (subtest 3 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to record responses verbatim	R	BP
4. Assigning an incorrect point value to a response	S	M
5. Inappropriate start point	A	M
6. Failure to establish basal	A	M

7. Failure to reach ceiling	A	M
8. Failure to stop subtest upon reaching ceiling	A	M
9. Failure to query the examinee when instructed by manual	A	M
18) Copying (subtest 8 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Assigning an incorrect point value to a response	S	M
4. Inappropriate start point	A	M
5. Failure to establish basal	A	M
6. Failure to reach ceiling	A	M
7. Failure to stop subtest upon reaching ceiling	A	M
8. Failure to keep drawings	A	BP
19) Early Number Concepts (subtest 10 on Early Years)		
1. Incorrect calculation of raw score	S	A
2. Failure to utilize an existing set	A	M
3. Failure to record responses verbatim	R	BP
4. Assigning an incorrect point value to a response	S	M
5. Inappropriate start point	A	M
6. Failure to establish basal	A	M
7. Failure to reach ceiling	A	M
8. Failure to stop subtest upon reaching ceiling	A	M
20) Matching Letter-Like Forms (subtest 11 on Early Years)		
1. Incorrect calculation of raw score	S	A

2. Failure to utilize an existing set	A	M
3. Failure to administer sample/practice/teaching items	A	M
4. Addition of scores for practice items in raw score	S	A
5. Failure to record responses verbatim	R	BP
6. Assigning an incorrect point value to a response	S	M
7. Inappropriate start point	A	M
8. Failure to establish basal	A	M
9. Failure to reach ceiling	A	M
10. Failure to stop subtest upon reaching ceiling	A	M

WJ III COG

Error description	Type	Source
1) Verbal Comprehension		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to establish basal	A	M
5. Failure to reach ceiling	A	M
6. Failure to stop subtest upon reaching ceiling	A	M
7. Failure to query the examinee when instructed by manual	A	M
8. Failure to administer full page when indicated by manual	A	M
2) Visual-Auditory Learning		

1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to stop subtest upon reaching ceiling	A	M
5. Failure to record time of day/date	R	M
6. Two aspects of one word (e.g. go...ing) are two errors, not one	S	M
7. Failure to indicate whether corrective feedback was given	A	M
3) Spatial Relations		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
4) Sound Blending		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
5) Concept Formation		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP

4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
6) Visual Matching		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to record completion time on timed subtests	R	M
7) Numbers Reversed		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to establish basal	A	M
5. Failure to reach ceiling	A	M
6. Failure to stop subtest upon reaching ceiling	A	M
7. Ceiling is 3 highest in a GROUP		
8) Incomplete Words		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
9) Auditory Working Memory		
1. Incorrect calculation of raw score	S	A

2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
10) Visual-Auditory Learning – Delayed		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to stop subtest upon reaching ceiling	A	M
5. Failure to record time of day/date	R	M
6. Two aspects of one word (e.g. go...ing) are two errors, not one	S	M
7. Failure to indicate whether corrective feedback was given	A	M
11) General Information		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to establish basal	A	M
5. Failure to reach ceiling	A	M
6. Failure to stop subtest upon reaching ceiling	A	M
7. Failure to query the examinee when instructed by manual	A	M
12) Retrieval Fluency		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M

13) Picture Recognition		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
14) Auditory Attention		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
15) Analysis-Synthesis		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to reach ceiling	A	M
5. Failure to stop subtest upon reaching ceiling	A	M
16) Decision Speed		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record completion time on timed subtest	R	M
17) Memory for Words		

1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to establish basal	A	M
5. Failure to reach ceiling	A	M
6. Failure to stop subtest upon reaching ceiling	A	M
18) Rapid Picture Naming		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record responses verbatim	R	BP
4. Failure to indicate time		
19) Planning		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Inappropriate start point	A	M
4. Calculation of correct responses instead of incorrect responses	A	M
20) Pair Cancellation		
1. Incorrect calculation of raw score	S	A
2. Assigning an incorrect point value to a response	S	M
3. Failure to record completion time on timed subtest	R	M