

WJ-III COG ADMINISTRATION ERRORS BY GRADUATE STUDENTS:
A VIDEO ANALYSIS

Luke W. Erichsen

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements for the degree

Doctor of Philosophy

in the School of Education

Indiana University

September, 2014

Doctoral Committee

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Jack Cummings, Ph.D., chair

Scott Bellini, Ph.D.

David Estell, Ph.D.

Rebecca Martínez, Ph.D.

August 5, 2014

Copyright © 2014

Luke Erichsen

Luke W. Erichsen

WJ-III COG ADMINISTRATION ERRORS BY GRADUATE STUDENTS:
A VIDEO ANALYSIS

Competent assessment practices allow psychologists to answer specific questions and formulate recommendations for intervention. One test, the Woodcock-Johnson Tests of Cognitive Abilities, 3rd Edition (WJ-III COG), has grown in popularity recently as a theory-driven test of cognitive skills. However, like all standardized tests, the WJ-III COG is only useful insofar as it is administered correctly. Minimizing error is crucial because results on standardized, norm-referenced tests only have meaning to the extent that they accurately measure the examinee's performance relative to the normative sample. Studies of other tests show that graduate students frequently commit errors, but previous research has only examined written products (test protocols) rather than assessing error on observed administrations.

This research analyzes data gathered as part of the normal training process for graduate students in school psychology. First-year students in a school psychology program were required to conduct practice administrations of cognitive tests with children and submit videotapes of these administrations to their course instructor. These materials were obtained for research purposes and analyzed for errors. This study represents the first systematic examination of administration errors on the WJ-III COG (or any other standardized intelligence test) employing analysis of videos. In total, 34 videos were analyzed from 15 examiners. An average of 34.5 errors were committed per video ($SD = 21.9$). All examiners committed errors, most frequently failing to read test directions verbatim and improper administration of corrective feedback procedures.

These errors have more potential to decrease examinee scores than the reverse. The vast majority of errors were not detectable by analysis of protocols alone.

Jack Cummings, Ph.D., chair

Scott Bellini, Ph.D.

David Estell, Ph.D.

Rebecca Martínez, Ph.D.

Acknowledgements

I am deeply thankful to my advisor, Dr. Jack Cummings, for his prolonged assistance throughout this project. Without his continued support for the research, frequent communication, and establishing of deadlines, I would not have succeeded. He displayed genuine enthusiasm for the research and its implications for the field that was infectious and motivating.

Thank you to the other members of my committee: Dr. Scott Bellini, Dr. David Estell, and Dr. Rebecca Martínez. Their feedback has been thoughtful and challenging, and my dissertation benefited immensely from their input. In addition, Dr. Julia Byers deserves special acknowledgement for her passion for teaching cognitive assessment and her role in facilitating recruitment for this study.

I would also like to thank fellow interns Josh Rainey, Jennie Purcell, and Janna Williams-Pitts for their support throughout the last year of this process, providing help with both data analysis and, more importantly, friendship.

Finally, I owe the greatest debt to my wife, Désirée Valentine, for her constant encouragement. Through many evenings and weekends spent researching, coding videos, and writing, she has been always supportive with kind words, motivation, and patience. To my daughter Evelyn, I both thank you for your bright spirit and apologize for the time away from you.

Table of Contents

Introduction.....	1
Examiner Error on the Wechsler Scales	3
Examiner Error on Other Tests	9
Statement of the Problem.....	13
Significance of the Problem.....	14
Research Questions.....	15
Review of the Literature	16
Assessment’s Continued Relevance	16
Assessment of Specific Learning Disabilities.....	17
The WJ-III COG and CHC Theory.....	19
Method.....	23
Participants.....	23
Procedure	26
Data Analysis.....	29
Results.....	33
Research Question 1	33
Research Question 2	41
Research Question 3	42
Discussion.....	43
Implications for Education and Practice	46
Limitations	51
Implications for Further Research	53
Conclusion	55
References.....	57
Appendix A: Adapted WJ-III COG Examiner Checklist.....	71
Appendix B: Approved Fall 2012 IRB Documentation.....	79
Appendix C: Approved Fall 2013 IRB Documentation.....	86
Appendix D: All Errors by Examiner and Administration.....	90
Appendix E: Rationales for Classification of Common Errors as Systematically Affecting Examinee Scores.....	105
Curriculum Vita	

List of Tables

Table 3.1: WJ-III COG Administration Materials Provided by Study Participants.....23

Table 4.1: Means and Standard Deviations of Errors across Three Test Administrations.....34

Table 4.2: Most Frequently Committed Errors on WJ-III COG Videos (>10 Errors).....35

Table 4.3: Number of Examiners Committing Error by Type with Potential Effect on Derived Scores (Errors Committed by >1 Examiner).....36

CHAPTER 1

Introduction

Over the last century, professional psychology has developed a rich tradition of research into assessment practices that has led to the development of highly sophisticated and psychometrically sound assessment instruments (Goldstein, 2013). Competent assessment practices allow psychologists to answer specific questions and formulate recommendations for intervention. Assessment instruments, such as intelligence tests, can aid the development of hypotheses when integrated with other data about an individual (Groth-Marnat, 2009). However, past research has shown that both graduate students (Ramos, Alfonso, & Schermerhorn, 2009) and practicing professionals (Brazelton et al., 2003) who administer intelligence tests often commit errors with alarming frequency. Examiner error limits the validity of conclusions based on test data and lowers trust in assessment.

Understanding the basic principles of human error can help make sense of why accomplished graduate students and practitioners commit errors even after formal preparation. Human error is a topic of interest in other fields as well, particularly within medicine, aviation, and engineering. In these fields, human error can result in major injury and loss of life. Reason's (1990) seminal *Human Error* is the most widely-cited text on the topic, and his more recent publications (e.g., Reason, 1997; Reason, 2000; Reason, 2008) apply the same model to reducing organizational and industrial accidents. Research interest in this area intensified after a series of high-profile disasters in the 1980s, including Three-Mile Island and Chernobyl, when human error and design issues contributed to nuclear accidents. The discipline of ergonomics (also called human factors)

owes a great deal to the work of Reason and his predecessors, and ergonomics has been applied within highly technical systems to reduce error (Woods et al., 2010). While applications of ergonomics are often extremely complex and far removed from research on examiner error as overviewed in the present study, knowledge of the basic concepts elucidated by Reason provides a foundation for appreciating that human error is both commonplace and understandable.

Reason (1990) is an academic cognitive psychologist and comes from an information processing theoretical framework. His research on human error unified previously separate lines of research that distinguished among *slips*, *lapses*, and *mistakes*. While Reason's model is much broader, these distinctions are most applicable to examiner error on tests. Slips, lapses, and mistakes are all types of error, defined as follows:

Error will be taken as a generic term to encompass all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when those failures cannot be attributed to the intervention of some chance agency. (p. 9)

As applied to examiner error, obtaining an accurate measurement of a child's cognitive abilities using a test's standardized procedure could be considered the "intended outcome." Error occurs when the planned sequence of activities (i.e., standardized administration) is not followed and the child's abilities are therefore not accurately measured. Reason (1990) noted that human error is inevitable and a byproduct of human cognition's extraordinary ability to simplify complex tasks by selecting, retrieving, and using stored knowledge structures in response to situational cues.

According to Reason (1990), the related concepts of slips and lapses occur during the *storage* and *execution* stages of a process. Slips are usually observable and happen when someone acts in a way that was not planned (regardless of whether a plan is adequate or not). Failing to correctly transfer raw score subtotals to a test's scoring software, or committing an arithmetic error, are examples of slips. The plan to follow these steps correctly was not executed properly. Lapses generally involve failure of memory—Reason (2008) provides the example of a nurse delivering a medication dose late. For examiners, failure to remember exactly what to say when a query is needed could constitute a lapse. Mistakes are more complex and may involve faulty plans that do not lead to an intended outcome, and they can involve a lack of knowledge or incorrect application of rules. Ambiguous scoring criteria or minor differences between versions of a test could lead to mistakes. Many significant errors are examples of *cognitive underspecification*, referring to applying a cognitive routine successful in one context to another similar but crucially different context, perhaps because that routine is frequently utilized. Mistakes, in particular, may be indicative of systemic factors involving preparation for conducting tasks (Reason, 2000). In the context of examiner error on tests, detailed knowledge about these errors can help ensure that the processes and methods used to prepare examiners to conduct assessments are effective.

Examiner Error on the Wechsler Scales

Decades of research have shown that examiners frequently commit a variety of errors on the most commonly used measures of intelligence, the Wechsler scales. These include the Wechsler Adult Intelligence Scale (WAIS), the Wechsler Intelligence Scale for Children (WISC), and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). As early as 1955, Plumb and Charles recognized that the Comprehension

verbal subtest on the original Wechsler-Bellevue test (the WAIS' immediate predecessor) was challenging to score, but fifteen years later Miller, Chansky, and Gredler (1970) still hypothesized that "WISC subtests lend themselves to highly objective scoring ... [interrater] ratings would be highly comparable" (p. 190). However, graduate student participants in the study who scored one identical protocol assigned FSIQ scores ranging from 76 to 93. The modal FSIQ, 80, was only agreed upon by 22% of the students. Multiple subsequent studies on the early Wechsler scales continued to disprove Miller et al.'s hypothesis that scoring was objective and reliable, and large discrepancies were frequently observed (e.g., Babad, Mann, & Mar-Haylm, 1975; Bradley, Hannah, & Lucas, 1980; Franklin, Stillman, Burpeau, & Sabers, 1982; Ryan, Prifitera, & Powers, 1983). Slate and Hunnicutt's (1988) review of the literature suggested that three factors were responsible for the majority of errors: inadequate instruction in cognitive assessment, ambiguity in test manuals, and examiner carelessness.

To address methodological problems and gaps in the literature, Slate and his colleagues conducted several well-designed studies of examiner error on the WAIS-R and WISC-R (Slate & Jones, 1990a; Slate & Jones, 1990b; Slate, Jones, & Murray, 1991; Slate, Jones, Coulter, & Covert, 1992; Slate & Jones, 1993). They used protocols from actual test administrations rather than contrived protocols and also systematically quantified the type of errors examiners committed; most previous researchers did not include details regarding particular subtests and items that were most challenging. This line of research intended to inform the development of better practices for teaching cognitive assessment courses, and subsequent studies generally follow their format and much of the methodology.

Slate and colleagues' research (Slate & Jones, 1990a; Slate & Jones, 1990b; Slate, Jones, & Murray, 1991; Slate, Jones, Coulter, & Covert, 1992; Slate & Jones, 1993) provided convincing evidence that examiner error was frequent on the WISC-R and WAIS-R and that these errors can ultimately affect placement decisions. Slate and colleagues distinguished between *recording errors* and *nonrecording errors*. Failure to record examinees' responses is a practice that does not necessarily directly impact scoring accuracy but does violate standardization and precludes both checking responses later and qualitative analysis of an examinee's errors (Slate, Jones, Murray, & Coulter, 1993). However, because recording errors (failure to record responses) are very common and acts of omission rather than commission, Slate and colleagues calculated error frequency both with and without nonrecording errors. More recent researchers followed suit. For the sake of clarity, I will follow Platt et al.'s (2007) example and instead use the alternate terms *commission errors* and *omission errors* rather than Slate's *nonrecording* and *recording* errors, respectively.

Slate and colleagues (Slate & Jones, 1990a; Slate & Jones, 1990b; Slate, Jones, & Murray, 1991; Slate, Jones, Coulter, & Covert, 1992; Slate & Jones, 1993) found that the typical scored WAIS-R or WISC-R contains a significant number of commission errors—varying from a mean of 8.7 errors per protocol to 16.9. Several studies included practitioners in addition to graduate students, and despite the more troubling consequences of error when working with actual clients, practitioners made at least the same number of commission errors as graduate students and more omission errors. For example, when 56 randomly selected WISC-R protocols were analyzed from a school system's records, the mean number of commission errors was 8.7 (Slate, Jones, Coulter,

& Covert, 1992). The nine practitioners who administered these tests reported administering, on average, 570 WISC-R's over their careers. Similarly, eight practitioners committed a mean number of 15.4 errors per protocol on the WAIS-R (Slate, Jones, Murray, & Coulter, 1993). Graduate students error rates were similar: 11.3 errors per protocol on the WISC-R, and 16.9 errors per protocol on the WAIS-R (Slate, Jones, & Murray, 1991; Slate & Jones, 1990). On the one existing study of a WPPSI scale, a small analysis ($n=57$ protocols) indicated a higher error mean rate of 27.1 commission errors per protocol on the WPPSI-R, perhaps due to the added difficulty of testing young children (Whitten, Slate, Jones, Shine, and Raggio, 1994).

Other notable findings from Slate and colleagues' research include details about the most common types of errors and the impact of errors on derived scores (Slate & Jones, 1990a; Slate & Jones, 1990b; Slate, Jones, & Murray, 1991; Slate, Jones, Coulter, & Covert, 1992; Slate & Jones, 1993). By far, the most errors occurred on the Verbal subtests (Similarities, Vocabulary, and Comprehension), and examiners tended to assign more points than were earned. This result suggests both ambiguity in the testing manual and difficulties in discerning between very similar responses that have different point values. Querying inappropriately (either querying when not needed or failing to query when required) was an especially common error. Other common errors included failure to use the correct basal or ceiling, totaling subtest scores incorrectly, and calculating chronological age incorrectly (these clerical errors, while simple, can have an especially damaging effect on derived scores). When protocols were rescored after corrections, the percentage of FSIQs affected by error ranged from 46% to 88%; examiners usually assigned a higher FSIQ than was earned. Across these studies, the majority of corrected

FSIQs were within two to three points of the original, but a significant percentage exceeded that figure. For example, one study included 27% of graduate student protocols having corrected FSIQs four or more points discrepant (Slate, Jones, & Murry, 1991).

Research on the next iteration of Weschler scales, the WISC-III and WAIS-III, indicated that examiner error continued to be highly problematic, although some small positive changes were evident (Alfonso, Johnson, Patinella, & Rader, 1998; Belk, LoBello, Ray, & Zachar, 2002; Brazelton et al., 2003; Platt et al., 2007; Ryan & Schnakenberg-Ott, 2003; Van Noord & Prevatt, 2002). The Vocabulary and Similarities subtests had somewhat reduced error rates, perhaps because the list of acceptable responses was moved from the appendix to a more logical location adjacent to the stimulus items in the administration manual (Alfonso et al.). FSIQ changed after corrections in 50% of protocols by an average of two points. Belk et al. found a lower average change in corrected FSIQ of about one point, but noted several extreme outliers resulting from mechanical or computational errors; e.g., 7% of protocols included FSIQ discrepancies exceeding 8 points. Ryan and Schnakenberg-Ott (2003) and Brazelton et al. (2003) compared practitioner error rates to graduate students using identical protocols and found fairly similar error rates, although Ryan and Schnakenberg-Ott's sample had significantly more variability in scores among students than psychologists. Brazelton et al. confirmed that among practitioners, work setting and degree (master's vs. doctorate, school vs. clinical) are insignificant factors contributing to error rates. Respondents who had administered more than 100 WISC-III's committed fewer errors than those who had administered less than 10.

Van Noord and Prevatt (2002) approached the topic of examiner error differently and argued that small modal differences between corrected and uncorrected scores are evidence of good psychometric properties; they wrote that “results of this study corroborate previous research findings of strong interrater reliability” (p. 174). Van Noord and Prevatt found a relatively small difference between corrected and uncorrected FSIQs of about one point; yet, their statement is surprising given previous research on examiner error. However, they acknowledged that results could be clinically significant without statistical significance. In their analysis, which included both the WISC-III and the Woodcock-Johnson Revised Tests of Achievement (WJ-R ACH), specific learning disability determinations using the discrepancy formula were changed for 2 out of 104 children in the sample. This finding is a reminder that changes of only a few points can mean the difference between artificial cutoff points as well as qualitative descriptors, such as *borderline* vs. *extremely low* or *high average* vs. *superior*.

Two studies of examiner error on the WISC-IV have been conducted to date (Loe, Kadlubek, & Marks, 2007; Mrazik et al., 2012). Both were similar in methodology and included a relatively small number of graduate student participants: 17 and 19, respectively. The authors of both studies drew similar conclusions from the data, that error rates on the fourth edition of the test are generally not improved from the WISC-III, largely because the Verbal Comprehension Index subtests are very similar to the previous version. Mrazik et al. observed a much lower frequency of “careless” computational or mechanical errors than in previously published research and noted that the course instructor penalized these types of errors on student grades. Because knowledge of

assessment course teaching methods informs intervention, variation in error rates based on pedagogical variables is another important yet understudied topic.

Examiner Error on Other Tests

Few researchers have devoted attention to intelligence tests other than the Wechsler scales (Hunnicut, Slate, Gamble, & Wheeler, 1990; Loe, 2014; Ramos, Alfonso, & Schermerhorn, 2009). One study on the Kaufman Assessment Battery for Children (K-ABC) was conducted when the test was newly published and novel in design compared to other batteries (Hunnicut et al.). Error rates were somewhat lower than on studies of the Wechsler scales, and the global score (Mental Processing Composite) was affected on only 35% of the 46 protocols analyzed to a small degree. However, Hunnicutt et al. noted that factors other than objectivity and administration ease may have accounted for this finding. In addition to this study of the K-ABC, two studies detailed errors on two versions of the Woodcock Johnson Tests of Achievement (Gurley, 2008; Van Noord & Prevatt, 2002). While not a measure of cognitive abilities and not comparable in content to the focus of the present study (the WJ-III COG), the WJ ACH's examiner manual, protocols, and the mechanics of scoring are similar in format to the cognitive test (Woodcock, McGrew, & Mather, 2001). Examiner error on both the WJ-R ACH and the WJ-III ACH was relatively low compared to the Wechsler scales except on the problematic Writing Samples subtest. Gurley found that failure to administer the entire page when required (a ceiling error) and incorrectly entering scores, grades, or ages into the computer software accounted for the majority of observed errors on the WJ-III ACH. Finally, Loe (2014) recently published an analysis of protocol errors on the Reynolds Intellectual Assessment Scales (RIAS), a brief measure of intelligence, and found that 90% of protocols contained commission errors.

There is the only published study of error on the WJ-III COG. For this reason, Ramos et al.'s (2009) analysis deserves additional attention. Thirty-six graduate students taking a cognitive assessment course from one instructor over three sections participated. They administered the full battery three times to volunteers and received feedback from the instructor throughout the learning process. Ramos et al. did not mention the use of videotapes or live supervision, although they did practice with classmates before working with volunteers. In total, 108 test records were included in the analysis, which involved examination by one of four advanced graduate students and scoring using Braden and Alfonso's (2002) examiner checklist. Interrater agreement was not calculated because Ramos et al. viewed the checklist as minimizing subjectivity.

Ramos et al. (2009) identified 500 errors in total, although 33% of these errors were made on five protocols and 46% of protocols only had zero or one error. In addition to frequent omission errors (failure to record incorrect responses and failure to circle the "total number correct" row), ceiling errors frequently occurred on four subtests. This type of error involves administering more items than is required or failing to administer items based on discontinue rules; violating basal and ceiling rules can affect scores. Of additional concern, Ramos et al. identified a total of 108 instances of students incorrectly entering raw scores into the computer software. They did not calculate the effect of incorrect raw score entry or other errors on derived scores, acknowledged as a limitation of the study. However, significantly fewer errors occurred on the third administration compared with the first (a decrease from 191 total errors to 133, $p < 0.05$).

Ramos et al. (2009) wrote that the large percentage (46%) of protocols with zero or one error "speaks to the ease of the administration procedures of the WJ-III COG" (p.

656). However, the types of error that can be detected based on examination of protocols alone is only one part—indeed, a small part—contributing to the overall ease of administration for a test. Only one study could be located concerning perceived test administration difficulty, and no version of the Woodcock-Johnson was included (Chattin & Bracken, 1989). However, when considered in the present context of examiner error, its results are concerning. Practicing school psychologists ($n=267$) identified a variety of characteristics that can contribute to intelligence test administration difficulty: organization of materials, organization of the manual, administration instructions, protocol layout, protocol scoring, manipulation of materials, and length of administration (Chattin & Bracken, 1989). The majority of school psychologists did not study three of four popular intelligence tests in graduate school, and a significant percentage reported feeling inadequately prepared to administer these three tests (although some respondents still did so). Additionally, although 100% of respondents felt prepared to administer the highly familiar WISC-R which 95% of the sample learned in graduate school, 12.5% reported sometimes having difficulty with administration instructions and 12.7% with manipulation of the materials. These results are deeply problematic and suggest with near certainty that administration errors were frequent, although many of these types of errors would not appear on a protocol.

Hopwood and Richard (2005) provided corroborating evidence that the type of errors found by examining protocols represents only a portion of error rates. Theirs is the only published study of examiner error to date that moves beyond having researchers or participants rescore protocols to study error rates. Rather than assessing error rates by rescoring completed protocols or asking participants to score examinee responses

recorded on unscored protocols, Hopwood and Richard asked two groups of graduate students to score the WAIS-III based on unscored protocols *or* film clips of a scripted actor providing the same responses as on the protocols. In comparison to students who scored the incomplete protocols, students who filled in and scored a blank protocol based on film clips committed more errors ($M=9.88$ errors per protocol on the video condition vs. 7.32 on the incomplete protocol condition). The associated effect size was large (Cohen's $d=.78$). Hopwood and Richard also supported their additional hypothesis that errors increase as a function of FSIQ; the two protocols used in this study had true scores of 85 and 112.

Hopwood and Richard's (2005) findings are especially noteworthy given that the film clip analysis more closely replicates scoring procedures in an actual administration than any previous study. They noted that the Verbal subtests require examiners to make some scoring judgments quickly during administration to know when to query or discontinue. Likewise, all subtests include basal and ceiling criteria that must be applied during administration. It is important to note that Hopwood and Richard's participants had the opportunity to replay clips for each test item, could pause the video, and were not interacting with an examinee. Hopwood and Richard noted that the most significant limitation of their study was that participants did not actually administer the WAIS-III: "A full WAIS-III administration would probably result in higher rates of scoring inaccuracy than we reported given the greater demands of an actual testing situation.... previous studies on scoring accuracy are likely to have considerably underestimated real-world errors" (p. 453). The demands of actual testing situations include working with the client to maintain rapport, motivation, appropriate behavior, and other demands (Sattler,

2008). Test manuals state that administration should be fluid and fast-paced to maintain examinee attention and reduce fatigue (Wechsler, 2003; Woodcock, McGrew, & Mather, 2007). Moreover, it is logical that testing with children, especially children with behavioral or attention problems, places additional demands on the examiner than testing the majority of adults on a WAIS.

Statement of the Problem

While Hopwood and Richard (2005) attempted to more closely replicate an actual testing situation by using film clips, they did not assess examiner *administration* error per se, but rather scoring error. Results on all extant studies of examiner error may be underestimates of error because examiners who originally completed the protocols could have committed errors that are not evident based on written records. Alfonso et al. (1998), Gurley (2008), and Hunnicutt et al. (1990) have also recognized that underestimates of error are likely given the lack of direct observation of examiner behavior.

Any number of errors could occur. At the most basic level, an examiner may misread the standardized directions or may make impermissible comments (e.g., answering questions when not allowed or telling an examinee a response was correct). An examiner may not provide required feedback or prompts on sample items. It is also the case that many details of administration are not recorded on protocols, particularly when examiners are required to present sample items using manipulatives or by pointing to items. The WJ-III COG includes several complex subtests that require the examinee to learn based on examiner prompts, pointing, and feedback, such as Visual-Auditory Learning and Concept Formation. A myriad of administration errors of various types and effects can occur throughout a test administration. A clear need exists for research that addresses this major gap in the literature: lack of direct observation of administrations. In

addition, only Ramos et al. (2009) have analyzed error on the WJ-III COG despite its increasing usage, and their study did not include an interrater agreement check.

Significance of the Problem

While the effect of error on derived scores is relatively low on average based on protocol analysis, particularly in more recent studies, a test's standard of error and the associated confidence interval only reflect statistical concerns such as sampling error and test-retest reliability; examiner bias represents another threat to validity beyond these factors (McDermott, Watkins, & Rhoad, 2014). All published studies include at least some protocols with a high number of errors and concomitant significant changes in scores. Importantly, even small changes can influence interpretation and thereby decisions about placement and intervention in educational settings. The problem of examiner error extends to settings serving adults as well, including tests administered for rehabilitation purposes and disability determination evaluations (Mpofu & Oakland, 2010). Of major concern, the issue of very small differences in derived scores has arisen in death penalty cases; for example, Florida recently attempted to execute a man before the U.S. Supreme Court decided in a 5-4 ruling that the state's use of a rigid IQ cutoff score of 70 for intellectual disability—and, therefore, potential eligibility for the death penalty based on the Court's *Atkins v. Virginia* (2002) ruling—was unconstitutional (Chappell, 2014; *Hall v. Florida*, 2014). The death row inmate had scored a 71 on the WAIS-III and a 72 on the WAIS-IV. The Supreme Court's decision will likely require states to follow current guidelines for determining the presence of intellectual disability rather than relying on strict cutoffs (Chappell, 2014), but standardized intelligence tests remain an essential component of DSM-5 intellectual disability diagnosis in addition to

assessment of adaptive functioning (American Psychiatric Association, 2013). As such, scores on tests can contribute to life-or-death decisions.

The recent popularity of cross-battery assessment, as discussed in the following chapter, also reemphasizes the importance of correct administration. School psychologists primarily use this method in the assessment of specific learning disabilities, which is the most common area of special education eligibility with over 2.3 million children receiving services for SLD (Data Accountability Center, 2011). Cross-battery assessments rely heavily on individual subtest scores across multiple batteries that all must be given correctly. Examiner error can more significantly affect results at this level, even if the change on a higher-level composite (such as FSIQ or a WJ-III Broad Cluster) is minor. Given the many variables involved in obtaining accurate scores, it is essential that psychologists eliminate the possibility of affecting decisions about examinees based on their own mistakes in test administration.

Research Questions

1. Based on analyses of videotaped WJ-III COG administrations and test protocols, how frequently do graduate students fail to follow standardized test administration procedures?
2. What effect do examiner scoring and administration errors have on derived WJ-III COG scores?
3. To what extent do errors observed on protocols alone differ from those observed by video analysis?

CHAPTER 2

Review of the Literature

The practice of standardized intellectual assessment is of continued relevance to the preparation and practice of school psychologists. In this chapter, I focus on the role of these assessment tools in the context of specific learning disabilities (SLD), an area in which their usage has grown increasingly controversial. The development and refinement of tests based heavily on modern intelligence theory has been concurrent with the move toward response-to-intervention (RTI)-based alternative service delivery models within schools, and I discuss how the Woodcock-Johnson Tests of Cognitive Abilities, 3rd Edition (WJ-III COG) plays an important role within these developments as some researchers advocate for the integration of RTI with the type of selective cognitive assessment facilitated by this testing battery in particular.

Assessment's Continued Relevance

Assessment constitutes a significant component of psychologists' graduate education, and intelligence testing is a fundamental component of such preparation. Despite the increasingly diverse environment in which professionals from a variety of disciplines deliver mental health services, psychologists continue to be the primary providers of formal assessment services (Krishnamurthy et al., 2004). Based on a 1999 survey of National Association of School Psychologist (NASP) members, school psychologists spend 46% of their time engaged in assessment activities, followed by consultation (16%; Bramlett et al., 2002). Castillo, Curtis, and Gelley's (2012) survey of NASP members indicates that little had changed by 2010: The number of special education evaluations conducted by school psychologists had steadily decreased over the

previous two decades, but school psychologists reported that they continue to spend nearly half their time (47%) conducting psychoeducational evaluations.

The role of school psychologists is evolving, but assessment continues to be highly valued by accrediting bodies and training directors. The Commission on Accreditation of the American Psychological Association requires all accredited doctoral programs to include preparation in assessment (APA, 2013), and NASP expects graduates of its accredited programs to be competent users of assessments as decision-making tools, including norm-referenced tests (NASP, 2010a). Doctoral internships in professional psychology also expect that students have developed competence in assessment. In a survey of Association of Psychology Postdoctoral and Internship Centers (APPIC) member sites, directors of clinical training reported that it was extremely important interns had extensive preparation in psychological assessment prior to internship, particularly in intelligence tests and objective personality measures (Stedman, Hatch, & Schoenfeld, 2001). Similarly, a survey of APA-accredited internship directors found that intelligence tests were the most frequently used assessment tools on internship, and over half of internships needed to offer instruction in basic introductory assessment methods including administration (Clemence & Handler, 2001). While these two studies are somewhat dated, the median number of integrated assessment reports completed by applicants to APPIC-member sites was the same in 2005 (7 adult reports, 5 child reports) as in the most recently available 2011 data (APPIC, 2011), suggesting that sites continue to value assessment.

Assessment of Specific Learning Disabilities

School psychologists are heavily involved in the identification of specific learning disabilities (SLD), a complex assessment task that has proven to be controversial. The

subsequent section explains how one method for assessing SLD relies on the cognitive theory that drives development of the WJ-III COG. SLD is the most common area of special education eligibility, and over 2.3 million children in the United States receive services under this category (Data Accountability Center, 2011). The move toward RTI models of service delivery has led some to deemphasize the role of intellectual assessment tools in the identification of learning disabilities (Gresham & Vellutino, 2010; Gresham, Restori, & Cook, 2008; Klassen, Neufeld, & Munro, 2005; Sternberg & Grigorenko, 2001). These researchers contend that careful evaluation of a child's response to evidence-based intervention over an appropriate period of time can provide adequate evidence for the presence of SLD when combined with information about contextual factors such as developmental history, behavior, and socioemotional considerations. However, others advocate that norm-referenced psychometric testing based on intelligence theory should supplement RTI-based methods for SLD identification, and that RTI's greatest utility is in prevention rather than identification (Flanagan, Ortiz, Alfonso, & Dynda, 2006; Flanagan, Fiorello, & Ortiz, 2010, Kavale & Spaulding, 2008).

This latter view was endorsed by the authors of a white paper developed in concert with the Learning Disabilities Association of America (Hale et al., 2010). Hale et al. argued that neither of the two methods of SLD identification allowed by federal law is sufficient alone—RTI or an ability-achievement discrepancy calculation. They agree with most others in the field that the discrepancy model is faulty. Instead, according to Hale et al., the most empirically-based approach—a “third method”—evaluates response to intervention within the context of a comprehensive evaluation assessing psychological

processing strengths and weaknesses using measures of cognitive abilities related to achievement (Hale et al.). Such an evaluation can potentially guide intervention based on correspondence between cognitive abilities and academic performance, although this research is young; for example, evidence-based instruction could be modified based on cognitive factors that moderate response to intervention (Fuchs et al., 2012).

Flanagan, Ortiz, and Alfonso (2008) and Hale et al. (2010) contend that current intelligence theory and well-designed tests allow for the identification of cognitive weaknesses empirically related to academic skills. Assessment allows educators to better understand and intervene when children do not succeed after receiving targeted evidence-based instruction (Tier II in an RTI model). These children have learning needs best met by individualized instruction informed by thorough assessment rather than simply more intensive Tier II interventions (Flanagan et al., 2008; Hale et al., 2010). Three prominent advocates of the “third method” approach to SLD identification wrote the following: “For the purpose of providing scientifically based intervention, RTI has no peer. For the purpose of providing scientifically based diagnostic information, cognitive assessment has no peer” (Flanagan et al., 2008, p. 17). While this controversy continues to develop within what has been called a paradigm shift (Jimerson, Burns, & VanDerHayden, 2007), it is likely that measures of cognitive abilities will continue to have a prominent place within SLD identification for the foreseeable future.

The WJ-III COG and CHC Theory

The Woodcock-Johnson Tests of Cognitive Abilities, currently in its third edition (WJ-III COG; Woodcock, McGrew, & Mather, 2007) is a particularly useful measure to guide SLD assessment. The WJ-III COG has become a frequently used measure of intelligence, particularly within school settings (Braden & Alfonso, 2003; Ramos,

Alfonso, & Schermerhorn, 2009). The measure has recently grown in popularity due to its theory-driven design and solid psychometric properties, as well as perceived ease of administration and computer-based scoring (Ramos et al.). Twenty years ago, 26% of school psychologists reported using the WJ-III COG's predecessor, the WJ-R COG (Stinnett, Havey, & Oehler-Stinnett, 1994). More recently, in university-based assessment training centers the WJ-III COG is utilized almost as frequently as the Wechsler scales, although the extent to which this usage transfers to practitioners is unknown (Orlovsky, Alfonso, & Kestenberg, 2005). Based on this finding from 2005, new graduates may be more likely to use the battery than experienced practitioners. Barak's (2008) analysis of assessment course syllabi from 71 school psychology programs indicated that 62% of instructors taught the WJ-III COG, exceeded only by the Wechsler Intelligence Scale for Children, 4th Edition (WISC-IV). These data indicate increasing usage of the test.

The WJ-III COG was designed to assess cognitive abilities as defined by the Cattell-Horn-Carroll (CHC) theory of cognitive abilities, a widely-recognized model based on decades of factor analytic research (Davidson & Kemp, 2011; McGrew, 2009; Willis, Dumont, & Kaufman, 2011). Nine broad abilities of intelligence are recognized and over 70 narrow abilities; the model continues to evolve. The two primary components of this theory, the Cattell-Horn *Gf-Gc* and Carroll Three Stratum models, were described by McGrew (2009) as "the consensus psychometric-based models for understanding the structure of human intelligence" (p. 1). Their synthesis as CHC theory has helped create a common language used by researchers and assessors of intelligence, particularly within school psychology but increasingly within other less applied research

disciplines (Davidson & Kemp, 2011; McGrew, 2009). For example, recent research utilizing CHC concepts in conjunction with the WJ-III COG has been published in *Child Development* (Hinnant, El-Sheikh, Keiley, & Buckhalt, 2013), *Journal of Abnormal Child Psychology* (McQuade et al., 2011), *Developmental Psychology* (Bub, Buckhalt, & El-Sheikh, 2011), *Personality and Individual Differences* (Jacobs, Szer, & Roodenburg, 2012), and *Intelligence* (Keith, Reynolds, Patel, & Ridley, 2008). In addition, the psychometric soundness of CHC theory as applied to testing batteries continues to be actively researched (e.g., Dombrowski, 2013; Reynolds, Keith, Flanagan, & Alfonso, 2013). Developers of the most recent versions of other intelligence tests besides the WJ-III COG—including the Stanford Binet, Kaufman Assessment Battery for Children, and the Differential Abilities Scale—all used CHC theory to help guide revisions and discussed these concepts in their technical manuals, and CHC is implicit in the most recent revisions of the Wechsler scales (Flanagan, Fiorello, & Ortiz, 2010).

Advocates of CHC theory's application to assessment argue that its sophistication allows for quality research on links between cognitive abilities and academic achievement that can guide both assessment and intervention (McGrew & Wendling, 2010). In this vein, a factor contributing to the WJ-III COG's popularity is its utility for cross-battery assessment, a method that has gained currency within school psychology as an evidence-based practice for identification of specific learning disabilities (Flanagan, Ortiz, & Alfonso, 2013). In cross-battery assessment, the examiner selects individual subtests from multiple assessment batteries that shed light on specific cognitive abilities relevant to the referral question. Selection of subtests is guided by CHC theory and research on which cognitive abilities are required for academic tasks, such as basic

reading skills. For example, children struggling with decoding and word recognition score lower on measures of the CHC constructs of comprehension-knowledge, long-term retrieval, processing speed, and short-term memory (McGrew & Wendling, 2010). Advocates of cross-battery assessment discourage the automatic tendency of many psychologists to give a full cognitive battery to every referred individual (Floyd, Keith, Taub, & McGrew, 2007). The WJ-III COG is popular among psychologists using cross-battery assessment because the test was designed based on CHC theory from its inception. As such, the subtests cover a wide range of cognitive tasks and also tend to load more “cleanly” within the underlying structure on specific cognitive abilities—i.e., factor analytic studies indicate that particular subtests within the WJ-III COG reliably measure specific cognitive abilities (Flanagan, Ortiz, & Alfonso, 2013; Reynolds, Keith, Flanagan, & Alfonso, 2013). However, like all standardized tests, the WJ-III COG is only useful insofar as it is administered correctly. Incorrect test administration by the examiner is unfair to the child and reduces the validity of the entire assessment process.

CHAPTER 3

Method**Participants**

School psychology graduate students taking a cognitive assessment course participated in the study. They were recruited from the Fall 2012 and Fall 2013 sections of P655, the cognitive assessment course at Indiana University. Students in P655 were required to submit three videotaped administrations of the WJ-III COG for course requirements. This test was the first taught in the class, although the third administration was completed at the end of the semester after students learned the WISC-IV. The final participation rate by students was 33% during the Fall 2012 semester: 9 students of 15 elected to participate, but only 5 provided their videos and protocols. Two video files from one participant were corrupt, and the participant did not possess usable files. In total, the Fall 2012 course yielded 13 usable videos across 3 administrations. I revised data collection procedures to increase response rate for the following year; 11 of 13 students elected to participate (85%), and 21 videos were obtained from these students. In total, 34 videos were analyzed for this study, including eight complete sets of three administrations. One student signed the informed consent form did not provide any videos or respond to follow-up requests. Table 3.1 indicates which test administrations were available for each participant.

Table 3.1

WJ-III COG Administration Materials Provided by Study Participants

Participant ID	Administration		
	1	2	3
1	X	X	X
2	*	*	X

Participant ID	Administration		
	1	2	3
3	X	X	X
4	X	X	X
5	X	X	X
6	X	X	X
7	X	X	X
8	X	X	X
9	X	--	--
10	X	--	X
11	X	X	X
12	X	--	--
13 (provided consent, but no videos)	--	--	--
14	--	--	X
15	X	--	X
16	X	--	X

Note: Two video files marked by asterisk were corrupt and unscorable.

Previous investigators (Alfonso et al., 1998; Loe et al., 2007; Ramos et al., 2009) identified the limitation of drawing participants from just a single program. To address this limitation and increase sample size, I sought participation from 12 other school psychology programs, focusing on other universities in the Midwest and on programs with faculty who have related research interests. Faculty members were contacted via email, provided with details about the study, and asked to forward an IRB-approved recruitment email to students in the cognitive assessment course, which included monetary incentives for both enrollment in the study and for providing all requested materials. While eight faculty members responded to email contact (66.7%), several (5) noted that they do not require videotaped administrations of the WJ-III COG, including two instructors who do not require video recordings of any tests. Two programs agreed to participate and forwarded recruitment information to students. However, despite two additional follow-up contacts with each course instructor, no students from other programs chose to participate in this study.

Instruments

Ramos et al.'s (2009) analysis of WJ-III COG protocols utilized an administration and scoring checklist available in Braden and Alfonso (2003) which was intended to account for all possible errors observable on protocols. However, the checklist's utility is clearly limited for the purposes of the present study. The *WJ-III Examiner Training Checklist*, available and reproducible from the *WJ-III Examiner Training Workbook* (Wendling & Mather, 2001), is better suited to investigating the research questions. This checklist includes many administration errors undetectable by protocol examination alone; for example, Test 2 (Visual-Auditory Learning): "Points immediately to the symbol and provides the word when subject makes an error" (p. 3). Some items, though, are challenging to quantify or are not actually required according to the examiner's manual, such as "Communicated to the subject that the session is enjoyable" (p. 1). Other possible errors mentioned in the Examiner's Manual or Standard Test Book are not included.

For these reasons, an adapted checklist was developed to meet the requirements of the present study and help account for additional potential errors (Appendix A). I developed these revisions to the *WJ-III Examiner Training Checklist* after careful examination of previous resources and the WJ-III COG test materials, as well as consideration of my experiences with supervision of first-year graduate students. A draft checklist was also reviewed by the P655 course instructor, a doctoral-level school psychologist with experience teaching and utilizing the WJ-III COG since its initial publication, and revisions were incorporated. However, unlike in studies of protocols, there is the potential for examiners to commit multiple and varied errors that can only be

detected by video analysis. The adapted checklist included space for the observer to document all errors with the minute and second that the error occurred in the video. Two additional columns allowed the observer to indicate if an error was likely to inflate or deflate the examinee's score. The original intent was to evaluate the possibility of score alterations on an item-by-item basis. However, it quickly became apparent during video coding that certain types of errors may change a score systematically in either a positive or negative manner. For this reason, I did not use these columns on the checklist and instead assigned categorizations only to error types for which I could articulate a clear rationale why that error would systematically affect raw scores across administrations.

Checklist items included were intended to be easily observable and unambiguous. The error that emerged as most frequently committed in the results, failing to read test directions verbatim, was operationalized as instances of an examiner inserting or substituting word(s) within text prompts printed in blue on the Standard Test Book (the only exception to this rule was made for failing to use the examinee's synonyms on Concept Formation). Instances when words were inserted before or after blue text prompts or when exact wording is not prescribed by the test were not coded as errors in this category, with the exception of instances when examiners posed queries as questions rather than requests ("Can you tell me another word?" rather than the correct prompt "Tell me another word."). Subtests 1-7 of the Standard WJ-III Battery were analyzed, as these are representative of the broad CHC factors and allow derivation of the General Intellectual Ability score.

Procedure

During the Fall 2012 semester, I first contacted the instructor of P655 at Indiana University to discuss recruiting students from her class and schedule a time to attend in

person. In consultation with the Indiana University Institutional Review Board, a procedure was developed addressing coercion concerns to ensure that neither the instructor nor other students would know who chose to participate in the study (see Appendix B for approved Fall 2012 IRB documentation). The course instructor explained the importance of the study while reminding students that nonparticipation would have no effect on their grade or whether they would receive corrective feedback. She then left the room, and I gave the students an informed consent form. After I answered questions, all students were asked to return the forms to me at the end of class, signed if they chose to participate. In total, 9 of 15 students agreed to participate and were asked to place their digital media and protocols in my mailbox after they had finished reviewing feedback from the course instructor. I sent reminder emails to the nine students after the instructor returned their final administration videos in December 2012 and another reminder in January 2013 when classes resumed. In total, five students submitted their materials. I uploaded the videos, along with scanned copies of the protocols, to Indiana University Box, a secure online storage service. In no point of the study process did I have access to the examinees' full names or birthdates, and ID numbers were used to identify the graduate student participants.

The preceding procedure was modified for recruitment during the Fall 2013 semester after an IRB amendment was approved (Appendix C). I again presented information about the study in person. Rather than rely on individuals who consented to participate to then provide me with materials, I asked the instructor to provide me with all students' materials after they were reviewed by her or advanced student supervisors. These materials were placed in sealed envelopes with names written on the outside.

Students who did not consent to study participation had their materials returned to their mailboxes unopened. This procedure maintained the privacy of students' choice to participate from the course instructor. The IRB amendment also permitted me to individually follow up with students in cases when materials were not initially available rather than sending group emails to all participants. In total, 21 videos from the Fall 2013 course were received from 10 students.

Scoring and interrater agreement. Videos of the three test administrations were originally assessed for accuracy by the course instructor (first set) and advanced graduate student supervisors (second and third sets). Protocols for all three administrations were reviewed by the instructor. The students received feedback on both protocols and videos. I scored all sets of videos myself using the adapted checklist developed for the present study. Eight videos (24% of the total sample) from eight different participants were independently scored by a doctoral candidate from an APA-accredited school psychology program for interrater agreement (IRA) purposes. One complete video administration was reviewed with the second rater before he began scoring.

The IRA analysis was based on error types (i.e., agreement if an error occurred in a video). To account for agreement that particular errors did not occur, the calculations included agreement that an error committed by at least 50% of examiners (Table 4.3) did *not* occur on the video. This adjustment, although somewhat arbitrary, reflects agreement beyond committal of errors and was also necessary for calculation of Cohen's kappa. Without such an adjustment, even "perfect" IRA for every committed error would result in a Cohen's kappa of 0. Because this statistic was designed to account for agreement by chance, it is heavily affected by a phenomenon's high base rate in the population; that is,

agreement would usually be expected by chance if the observed phenomenon is very common (Viera & Garrett, 2005). Cohen's kappa is typically used when two raters are making a judgment about an ambiguous observation, for example, in determining how reliably a structured interview detects the presence of a mental disorder when used by two raters, or whether a tumor is malignant or benign. The concept of "agreement by chance" assumes that raters sometimes simply guess when they do not know how to make a difficult categorization (McHugh, 2012). The determinations to be made in my study were largely objective. Disagreement was likely due to reasons other than guessing, perhaps a lack of sufficient review of the test and study procedures, and the kappa statistic should be interpreted with caution.

The initial IRA calculation indicated moderate agreement: 75.9% agreement on observations ($\kappa = .522$). Of the 26 disagreements, 23 occurred when I identified an error that the second rater did not. A third rater, a doctoral candidate from an APA-accredited clinical psychology program, reviewed video clips for 10 randomly selected disagreements and the associated administration rule. In every case, she agreed with my determination.

Data Analysis

Research Question #1: Based on analyses of videotaped WJ-III COG administrations and test protocols, how frequently do graduate students fail to follow standardized test administration procedures?

After scoring all videos using the adapted study checklist, data were entered into a Microsoft Excel spreadsheet. Descriptive statistics were then calculated using standard Excel functions and the Excel Data Analysis Toolpak. To address Research Question #1, I provide a breakdown of errors by type and per subtest with measures of central

tendency and ranges, as well as a total number of errors per administration. My presentation of descriptive data follows the models of Loe et. al (2007) and Mrazik et al. (2010) as adapted to the needs of the present study. Chapter 4 includes tables disaggregating data in several ways, including total number of errors across the three test administrations, most frequently committed errors, and a breakdown of error types by percentage of examiners committing the error.

Research Question #2: What effect do examiner scoring and administration errors have on derived WJ-III COG scores?

Some previous studies of examiner error on standardized tests include calculations of the effect of scoring errors on derived scores. For example, incorrectly summed raw scores can be correctly added, and then derived scores properly calculated. Likewise, if scores below a basal are not counted, the effect of correctly including the basal points will demonstrate how much the resulting score will increase. However, video analysis is fundamentally different in that the precise effect of most errors cannot be determined. At best, an estimate can be provided of which errors are likely to contribute to altered scores. As discussed previously in this chapter, I assigned categorizations to error types for which I could articulate a clear rationale why that error would systematically affect raw scores across administrations (Appendix E). Chapter 4 includes calculations of the percentage of errors with potential to increase or decrease scores.

Research Question #3: To what extent do errors observed on protocols alone differ from those observed by video analysis?

A brief comparison with Ramos, Alfonso, and Schermerhorn's (2009) analysis is needed to help explain the methodology pertaining to this research question. According to their study, the WJ-III COG compares favorably to the Wechsler scales in that relatively few errors are observed based on analysis of protocols alone. Ramos et al. observed no more than one error on 46% of the 108 protocols in their sample, and two of the three most common errors they observed have no direct impact on scores. These two error types were "failure to record examinee errors" and "failure to encircle correct row." Recording examinee errors is recommended by the Examiner's Manual but not mandated: "When possible, record incorrect responses verbatim on the Test Record for diagnostic purposes" (p. 35). Likewise, "encircling the correct row" simply provides the examiner with an age- and grade-equivalent estimate of raw scores without using the computer scoring software. The cognitive assessment course instructor in the current study did not require students to record examinee errors or encircle the correct row, and many students did not consistently do so; these were not considered errors. The cover page of the protocol (Identifying Information and Test Session Observations Checklist) was not analyzed.

All protocols analyzed in this study were initially scored by the course instructor for teaching and grading purposes. I rescored half (17) of the protocols for errors and did not observe any instances when the course instructor failed to notice a protocol error. These include basal and ceiling errors, arithmetic errors, failure to record administration of sample items, failure to record the exact time of Visual-Auditory Learning administration, and incorrect scoring of verbal responses if recorded. The course

instructor's examination of protocol errors are presented descriptively and compared with my analysis of video errors in Chapter 4.

CHAPTER 4

Results

This chapter presents analyses of errors on 34 videotaped administrations of the WJ-III COG from 15 participants. These analyses address the research questions described in Chapter 1 and include descriptive statistics of the number and types of errors committed by examiners, an estimate of the effect of errors on examinee scores, and a comparison of results with what would be obtained from only analyzing protocols rather than videos. Videos were scored for errors using a checklist developed for this study adapted from the *WJ-III Examiner Training Checklist* (Wendling & Mather, 2001). Results from the checklists were transferred to a Microsoft Excel spreadsheet, and analyses were conducted within Excel.

Research Question 1: *Based on analyses of videotaped WJ-III COG administrations and test protocols, how frequently do graduate students fail to follow standardized test administration procedures?*

Graduate student examiners committed an average of 34.5 errors per video ($SD = 21.9$). Across all 34 videos, examiners committed 1,173 errors. These statistics were affected in particular by three videos from three different examiners. These three videos were recorded for two second administrations and one first administration; with these data points removed, $M = 29.5$ and $SD = 15.1$. However, these three data points represent valid observations of problematic test administrations, and none were extreme enough to exceed a z -score of 3. As such, they were not considered outliers and are included in this analysis. A complete breakdown of errors for each study participant is included in Appendix D.

The most frequently committed error was failing to give test directions verbatim. This was the only error committed on every video ($M = 13.5$ errors per video, $SD = 10.2$, range 1-41). This type of error ranged from paraphrasing directions completely (for example, saying *what kind of clothes is he wearing?* rather than *what kind of garment is this?*) to using the wrong article on Concept Formation items (for example, saying *tell me the rule for a drawing to be inside the box* rather than *inside a box*). As verbatim errors accounted for 30% of all errors, error rates were also calculated without this error type included. Table 4.1 summarizes error frequency for each of the three administrations, both with and without verbatim errors.

Table 4.1

Means and Standard Deviations of Errors across Three Test Administrations

	All Errors		Errors (Not Incl. Verbatim)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
All Videos ($n = 34$)	34.5	21.9	21.0	16.4
First ($n = 13$)	36.9	22.6	22.8	14.7
Second ($n = 8$)	47.6	25.9	29.3	18.2
Third ($n = 13$)	24.0	13.4	14.0	7.4

Note: Second administration included two of three videos with unusually high number of errors.

A small number of error types accounted for a high proportion of total errors: The top five most frequently committed errors represent 72% of all errors committed (see Table 4.2). However, an “error” as a unit type may not be an especially helpful measurement for practical purposes for two primary reasons. First, errors are not created equal; for example, failing to give sample items has more potential impact on an

examinee's scores than waiting seven seconds to correct a Visual-Auditory Learning error rather than five. Second, some examiners committed few errors in total but nonetheless committed an average number of error *types*. That is, they committed multiple errors but only did so few times per administration. As examiner error on corrective feedback procedures was especially prevalent, examiners had more opportunity to commit errors when examinees committed more errors.

Table 4.2

Most Frequently Committed Errors on WJ-III COG Videos (>10 Errors)

	Number of Errors (<i>n</i> = 34)	Range Per Video
Global: Failed to read test directions verbatim	460	1-41
Visual-Auditory Learning: Failed to point to symbols when correcting	138	0-23
Visual-Auditory Learning: Too long before providing word on errors (>6 seconds)	87	0-18
Global: Gave examinee inappropriate feedback	80	0-21
Global: Recorded examinee response inaccurately on protocol	77	0-10
Concept Formation: Failed to acknowledge correct responses (to #35)	39	0-7
Concept Formation: Failed to use examinee's synonyms on feedback	31	0-6
Concept Formation) Failed to query responses appropriately	23	0-9
Concept Formation) Failed to provide corrective feedback on errors (to #35)	20	0-7
Numbers Reversed: Paused if examinee needs more time	17	0-3
Numbers Reversed: Presented oral item prompts too fast or slow	16	0-7
Verbal Comprehension: Failed to query verbal response as instructed	15	0-2
Visual-Auditory Learning: Mispronounced stimulus word (including <i>the</i> , suffixes)	13	0-2
Visual Matching: Failed to hold up book after sample on VM2	13	0-1

Table 4.3 presents data in an alternative format, detailing the number of examiners committing each error on at least one video regardless of how many times the error was committed, sorted by frequency within each subtest. All errors committed by at least two examiners are included. These data allow for better understanding of which subtests and specific errors were problematic for the most examiners. In this narrative, I will review frequently committed errors; all percentages refer to the total number of participants ($n=15$) and indicate the percentage of examiners who committed an error type on at least one video. The most problematic subtest was Test 2 (Visual-Auditory Learning); all examiners committed at least one error on this subtest. This is especially problematic given the nature of the test. As noted in the Standard Test Book, “This test is a controlled learning task. For this reason, all subjects must have an *identical* opportunity to learn” (Woodcock, McGrew, & Mather, 2001, p. 63, italics in original). Most frequently, examiners allowed too much time before providing words (13, 86.7%) or did not point to the symbol when providing its corresponding word (10, 66.7%). Some examiners (5, 33.3%) improperly pronounced stimulus words when introducing symbols, often by spelling out Story 7 suffixes letter by letter (*sss* and *ing*) or introducing *the* as *thē* even though *the* never precedes a word beginning with a vowel sound on any story. This error appeared to cause confusion for multiple examinees.

Table 4.3

Number of Examiners Committing Error by Type with Potential Effect on Derived Scores (Errors Committed by >1 Examiner)

	<i>n</i>	Potential Effect	%
Global: Failed to read test directions verbatim	15	---	100
Global: Gave examinee inappropriate feedback	11	---	73.3

	<i>n</i>	Potential Effect	%
Global: Recorded examinee response inaccurately on protocol	13	---	86.7
Test 1: Verbal Comprehension	13		86.7
Failed to query verbal response as instructed	8	Decrease	53.3
Failed to request 1-word responses	6	Decrease	40.0
Failed to establish ceiling	4	Decrease	26.7
Mispronounced stimulus word	3	Decrease	20.0
Failed to administer sample item	3	Decrease	20.0
Queried an entirely incorrect response	3	Increase	20.0
Failed to immed. reverse at end of page for basal	2	---	13.3
Antonyms: Failed to ask for response if “non”/”un”	2	Decrease	13.3
Failed to follow procedure if sample incorrect	2	Decrease	13.3
Failed to establish basal	2	Increase	13.3
Test 2: Visual-Auditory Learning	15		100
Too long before providing word on errors (>6s)	13	---	86.7
Failed to point to symbols when correcting	10	Decrease	66.7
Mispronounced stimulus word (e.g. <i>the</i> , suffixes)	5	Decrease	33.3
Provided second symbol (suffix) on Story 7	5	---	33.3
Failed to query skipped symbols	3	Decrease	20.0
Failed to correct an error	2	Decrease	13.3
Failed to count extra words as errors	2	Increase	13.3
Allowed examinee practice/time on samples	2	Increase	13.3
Test 3: Spatial Relations	6		40.0
Failed to give corrective feedback on sample items	4	Decrease	26.7
Test 4: Sound Blending	10		66.7
Failed to allow adequate time for responses	4	Decrease	26.7
Miscued audio track	2	---	13.3
Mispronounced stimulus word (in sample)	2	Decrease	13.3
Test 5: Concept Formation	15		100
Failed to acknowledge correct responses (to #35)	9	Decrease	60.0
Failed to query responses appropriately	9	---	60.0
Failed to use examinee’s synonyms on feedback	8	Decrease	53.3
Failed to provide feedback on errors (to #35)	6	Decrease	40.0
Improperly corrected an examinee’s synonym	2	Decrease	13.3
Failed to ask examinee to repeat sample correctly	2	Decrease	13.3
Test 6: Visual Matching	11		73.3
VM2: Failed to hold up book after sample	11	Increase	73.3
Test 7: Numbers Reversed	12		80.0
Paused if examinee needs more time	5	Decrease	33.3
Gave example item for incorrect start point	5	---	33.3
Presented oral item prompts too fast or slow	4	---	26.7
Presented item backwards	3	---	20.0
Ceiling error	3	Decrease	20.0

	<i>n</i>	Potential Effect	%
Cued audio track incorrectly	2	---	13.3
Prompted examinee to finish example item	2	---	13.3
Failed to give sample item after reversal for basal	2	Decrease	13.3
Basal error	2	Increase	13.3

Test 5 (Concept Formation), the other subtest with a significant controlled learning aspect, was also problematic with at least one error committed by all examiners. Most frequently, examiners failed to acknowledge every correct response through Item 35 (9, 60%) and failed to query responses appropriately (9, 60%). Nearly as many examiners (8, 53.3%) failed to use the examinees' synonyms when providing corrective feedback. For example, if an examinee incorrectly responds that the rule to a puzzle is "circular and red," the examiner should provide feedback by saying "The answer is 'circular and yellow'" even though the examiner prompt would in this case read "The answer is 'round and yellow.'" This is an especially noteworthy error because examiners are only instructed on this administration rule in the Examiner's Manual, *not* the Standard Test Book used during administration; no reminder of this administration rule is found in the Standard Test Book. Another frequently committed error on this subtest was failing to provide corrective feedback on errors through Item 35 (6, 40%).

Test 1 (Verbal Comprehension) also proved difficult for a significant portion of examiners, with 86.7% of examiners (13) committing an error. Most frequently, verbal responses were not queried as instructed (8, 53.3%), and one-word answers were not requested when the examinee provided a response with two or more words (6, 40%). While comparable errors on Wechsler verbal tests occur much more frequently based on protocol analysis (Loe et al., 2007), these types of errors on the WJ-III COG can be more

difficult to detect if the examiner does not record incorrect examinee responses, which was the case for many of these errors. Some examiners failed to establish a ceiling (4, 26.7%). Three other errors were committed by three examiners each (20%): mispronouncing stimulus words, typically advanced vocabulary words; failing to administer a sample item; and querying a completely incorrect response.

Most examiners (12, 80%) committed at least one error on Test 7 (Numbers Reversed), although this subtest is ostensibly a simple task to administer. Most frequently, examiners did not pause the audio track when the examinee was still in the process of answering (5, 33.3%) or gave unnecessary example items associated with the wrong starting point (5, 33.3%). Several examiners presented orally administered items markedly too fast or slow (4, 26.7%), and three examiners (20%) presented items backwards or committed a ceiling error.

The remaining subtests were less problematic and typically only contained one or two frequently observed errors. On Test 3 (Spatial Relations), 40% (6) of examiners committed at least one error, with failing to give corrective feedback on sample items the most frequently observed error (26.7%). Test 4 (Sound Blending) videos contained errors for 60% (8) of examiners, most frequently not pausing the audio track when the examinee needed more time (4, 26.7%). Finally, on Test 6 (Visual Matching), 73.3% (11) of examiners committed an error, usually failing to hold up the protocol after giving the sample items on VM2 to prevent the examinee from studying items before timer begins.

Correlations between Age and Errors (Posthoc Analysis) Many examiner errors occurred when the examiner was required to respond to an examinee's error. These include the frequently committed errors of failing to query verbal responses, properly

correcting errors on Visual-Auditory Learning, and providing corrective feedback on Concept Formation. When examinees struggled with these subtests and the examiner was required to provide extensive feedback, error rates were often high. Conversely, high-performing examinees allowed examiners few opportunities to commit certain errors. A pattern emerged during data analysis suggesting that a direct relationship exists between *examinee* error rates and *examiner* error rates, particularly on these corrective feedback subtests. To test this posthoc hypothesis, bivariate Pearson product moment correlations were calculated between examinee age (as reported on each protocol) and error total, age and error total without verbatim errors, and age and error type count. While a more direct test of this hypothesis would assess the correlation between examinee raw score and examiner error, age has more practical implications.

Age and error total was moderately negatively correlated, $r(32) = -0.56$, $p < .01$. Age and error total without verbatim errors was also negatively correlated, $r(32) = -0.47$, $p < .01$. Age and error type count were not correlated at a significant level, $r(32) = -0.26$, $p = .14$. This finding indicates that the significant relationship lays in the difference between these two calculation methods; i.e., examiners tend to commit certain errors more frequently with younger examinees, but the relationship between age and number of error types is not supported. However, the strength of association between total error rates with age is robust and has implications for graduate training, to be discussed in Chapter 5.

Research Question 2: *What effect do examiner scoring and administration errors have on derived WJ-III COG scores?*

Of the 41 error types listed on Table 4.3, which delineates errors committed by at least two examiners, 28 have an associated rationale for a systematic potential effect on examinee scores (Appendix E). The great majority of these error types (22, 78.6%) have the potential to decrease examinee scores. Less than one-quarter (6, 21.4%) are more likely to increase examinee scores. Of the error types with potential to increase scores, only two were committed by more than two examiners: failure to hold up the protocol after sample items on Visual Matching 2 (73.3% of examiners) and querying an entirely incorrect response on Verbal Comprehension (20% of examiners). In contrast, of the 22 error types with potential to decrease examinee scores, 15 were committed by at least two examiners. Three error types with potential to decrease examinee scores were committed by more than half of examiners: failure to point to symbols when correcting on Visual-Auditory Learning, failure to query verbal responses as instructed on Verbal Comprehension, and failure to use examinee's synonyms when providing feedback on Concept Formation.

Of the 14 errors committed most frequently (at least 10 times in total, see Table 4.2), 8 are classified as potentially decreasing examinee scores, 1 as potentially increasing examinee scores, and the remainder have an indeterminate effect. In total, 296 errors represented on Table 4.2 are classified as potentially decreasing examinee scores (25.2% of all errors, or 41.5% of all non-verbatim errors). The single error type represented on Table 4.2 that has potential to increase examinee scores, failing to hold up the book after Visual Matching 2's sample items, is only possible once per administration

and occurred 13 times in the sample, with 11 examiners committing the error on at least one administration.

Research Question 3: *To what extent do errors observed on protocols alone differ from those observed by video analysis?*

Consistent with the previously published study of the WJ-III COG (Ramos et al., 2009), relatively few errors are observable without viewing examiners giving the test. Across 34 test protocols in total, the course instructor identified 31 errors that do not rely on video observation. The modal number of errors per protocol was zero (15 protocols), followed by one error (12 protocols). Two protocols contained two errors each, while one protocol each contained three, four, or five errors respectively. Of the 31 total errors, 32% (10) were noticed by examiners before turning in the administration for grading (students were not penalized by the instructor for noting errors themselves before submitting materials, providing an incentive to recheck scoring).

The most common error on protocols was failing to record if a sample item was given (10 errors), followed by failure to establish a ceiling (7) and failure to establish a basal (6). Examiners also tested beyond the ceiling (3), incorrectly summed raw scores (2), failed to write the exact time of Visual-Auditory Learning administration (2), and scored a verbal response incorrectly (1).

This protocol-only analysis stands in contrast to findings obtained from video analyses. The total number of errors observable on protocols, 31, is 2.6% of the total number of errors in this sample, 1,173. Every video administration contained at least one error, while 43% (10) of the protocols were error-free. All 14 of the most frequently committed errors (Table 4.2) are unobservable on protocols.

CHAPTER 5

Discussion

This study is the first systematic examination of administration errors on the WJ-III COG (or any other standardized intelligence test) employing analysis of videos. Using a scoring checklist adapted from the test training materials, 34 videos were analyzed from 15 graduate student examiners. All examiners committed errors, most frequently failing to read test directions verbatim and improper administration of corrective feedback procedures. These errors have more potential to decrease examinee scores than the reverse. The vast majority of errors were not detectable by analysis of protocols alone.

These findings support previous research on test protocols indicating that graduate students frequently commit errors (Loe et al., 2007; Mrazik et al., 2010; Platt et al., 2007). These most recent studies of Wechsler scales indicate that virtually all examiners commit errors. Likewise, every examiner in my study committed errors, with a mean of 34.5 errors per video ($SD = 21.9$). My study significantly extends the existing research and indicates that the extant literature severely underestimates the prevalence of errors. The findings support the opinion of several previous researchers who hypothesized that studies using only test protocols underestimate error rates (Alfonso et al., 1998; Gurley, 2008; Hopwood & Richard, 2005). The most recent published study of examiner error, an analysis of the Reynolds Intellectual Assessment Scales, succinctly makes this claim and encourages future research in the vein of my study:

The use of protocol review to evaluate errors, though common throughout the published research in this area, cannot accurately reflect the true frequency and

impact of administration and computation errors.... Future research should incorporate videotaped test administration in an attempt to obtain a more accurate accounting of examiner errors (Loe, 2014, p. 105).

Using Reason's (1990) framework of human error, most errors observed in the current study are best categorized as *slips*, i.e., observable errors occurring when examiners fail to follow a predetermined plan. A few errors, however, are more attributable to a faulty plan (*mistakes*) and could be remedied by publisher revisions, such as more obvious instruction to use examinees' synonyms during Concept Formation; recall that the associated rule does not appear in the Test Book used during administration.

A comparison with the only published analysis of error on the WJ-III COG (Ramos et al., 2009) illustrates the extent to which current literature underestimates examiner error. Ramos et al. observed significantly lower error rates on WJ-III COG protocols than had been observed on Wechsler tests by previous researchers. Notably, nearly half (46%) of the 108 protocols in their sample contained either 0 errors or 1 error, although five protocols contained a very high number of errors. Ramos et al.'s participants administered 14 of the WJ-III COG's 20 subtests, and the authors employed a highly conservative measure of error; two of the three most commonly committed errors involved recording and have no direct effect on examinee scores. As such, their figures are not directly comparable with protocol analysis of my study, as these particular errors were not considered as such. Ramos et al. contended that the relatively low number of errors observed in their study attested to the WJ-III COG's ease of administration. Overall, my findings support the notion that relatively few errors are committed on protocols. However, the fact that only 2.6% of the total number of errors

were observable on protocols suggests that standardized administration of this test is, in fact, difficult for many examiners.

My findings further extend Hopwood and Richard's (2005) unique study using protocols and video clips associated with a scripted WAIS-III examination. They found that the increased task complexity associated with scoring a protocol based on a video of a person providing responses resulted in higher error rates than when examiners were asked only to score an incomplete protocol recording the same responses. Examiners even had the benefit of replaying video clips for each test item and pausing the video. The most significant limitation of their study, as noted by Hopwood and Richard, was that participants did not actually administer the test. My findings confirm their argument that the demands associated with test administration have been underestimated in previous research. Hopwood and Richard also found that errors increased as a function of FSIQ because examiners had more opportunity to commit errors on non-zero responses (particularly verbal items). My findings suggest that error rates on the WJ-III COG are higher when testing examinees who score lower; this is likely because many of the frequently committed errors involved corrective feedback procedures (7 of the 14 error types committed more than 10 times in total), which are not incorporated frequently into the Wechsler scales. The most common error on Wechsler scales across multiple studies involves assigning too many points to verbal responses and results in inflated derived scores, whereas errors on WJ-III COG corrective feedback procedures—particularly the Visual-Auditory Learning and Concept Formation controlled learning tasks—likely produce deflated derived scores.

Implications for Education and Practice

These results indicate that the WJ-III COG is a challenging test to administer exactly following standardized procedures. NASP (2010b) recognizes this as an ethical issue, requiring in the *Principles for Professional Ethics* that “when using standardized measures, school psychologists adhere to the procedures for administration of the instrument that are provided by the author or publisher or the instrument [or noting otherwise in the report]” (p. 7). Cognitive assessment course instructors hold the primary responsibility for educating new graduate students in meeting this ethical obligation to administer tests correctly and can benefit from a research-based understanding of how to instruct students in test administration.

The current study provides valuable information regarding which portions of the WJ-III COG are most challenging for examiners. It is recommended that instructors provide additional explicit instruction during class time on the most frequently observed errors based on Tables 4.2 and 4.3. Little is known about the most effective means of teaching test administration, and most investigations of this topic are dated. One quasi-experimental study of the WISC-R suggested that targeted instruction addressing commonly committed errors can reduce error rates compared to a group not receiving such instruction (Slate & Jones, 1989). The experimental intervention consisted of an additional 2-hour lecture discussing test procedures problematic for the control group with associated strategies for avoiding these errors. McQueen et al. (1994) also observed improvements in WISC-R error rates using a laboratory component taught by teaching assistants that gave additional attention to common errors as determined by previous research. Therefore, instructors of the WJ-III COG should consider providing systematic

instruction on common errors, based on results of the current study, in addition to individualized feedback.

Low sample size prevented analysis of an original intended research question, whether error rates improved across the three videotaped administrations. Simple practice is the only specific instructional technique that researchers have explored frequently; repeated student administrations of tests are easily studied as these are typically the most heavily weighted assignment in assessment courses (Barak, 2008; Cody & Prieto, 2000). Studies of Wechsler protocols indicate that repeated practice does not appreciably reduce error rates (Belk et al., 2002; Loe et al., 2007; Mrazik et al., 2012). Platt et al. (2007), however, observed modest improvement with practice and made a case for methodological issues as contributing to previous studies showing no significant improvement, specifically a failure to control for IQ of examinee volunteers and a lack of assurance that corrective feedback and instruction was provided before each successive administration. Moreover, Ramos et al. (2009) did observe lower error rates across administrations of the WJ-III COG. Unfortunately, my study does not help to answer this important question.

Egan, McCabe, Semenchuck, and Butler (2003) proposed a simple intervention for reducing errors that could easily be applied to courses utilizing videotaped administration and feedback. Their control group received instruction similar in format to the students in my study: lecture and demonstration of proper administration, followed by practice administrations outside of class and corrective feedback on errors after each administration. Participants in their experimental group placed completed protocols and written feedback in a portfolio; on each successive administration, they were asked to

review the materials and turn in the entire portfolio each time. This portfolio served as a personalized reference file with examples of proper scoring as well as how to correct errors. Great improvements were observed across five administrations compared to a control group. Egan et al. recommended also requiring students to keep a running list of their specific errors on the inside cover of the portfolio. This instructional method could be incorporated into courses that provide feedback on videotaped administrations by requiring students to maintain such a portfolio that includes feedback from reviewed videos.

Visual reminders in the Test Book and/or protocols based on my results could help students during the learning process. Students could be asked to prepare Post-It notes with reminders about frequently committed errors and place them on appropriate Test Book pages prior to administration, or write reminders on protocols before giving the test. A few protocols in the sample had such reminders, for example, one student wrote “Point to symbols” on the Visual-Auditory Learning page, and another wrote “Acknowledge right answers” on the Concept Formation page. This intervention would also require students to further immerse themselves in the test materials prior to test administration.

Importance of Requiring Videotaped Administration. The vast majority of errors cannot be detected by analysis of protocols alone; only 2.6% of total errors were observable on protocols. It is likely that a high proportion of errors on other tests are also only observable via video analysis. While some popular tests incorporate corrective feedback less frequently, other tests introduce additional opportunities to commit errors through their use of manipulatives, stimulus books, and additional test materials. For

example, all three current (4th edition) Wechsler scales include the Block Design subtest and subtests requiring a separate examinee response booklet. The SB5 and DAS-II incorporate manipulatives extensively. The WJ-III COG, presented using only the Test Book, is relatively streamlined in comparison.

Because most errors are not detectable on WJ-III COG protocols, observation of test administration is essential to provide comprehensive feedback to students. Students also stand to benefit greatly from reviewing their own recorded administrations as part of the learning process; instructors should consider requiring students to do so. It is unclear how frequently videotapes or live administrations are utilized in cognitive assessment courses. An analysis of course syllabi from 71 school psychology programs indicated that 60% of instructors required students to pass a “competency exam” to assess test administration (Barak, 2008). Of the instructors who gave competency exams, 40% used live observation of administration, while 20% required students to submit videotapes. An earlier study of cognitive assessment courses (Cody & Prieto, 2000) did not provide percentages but listed the average number of videotapes required by 94 course instructors ($M = 1.8$, $SD = 3.0$); however, video recording technology was perhaps not as accessible when this earlier study was conducted. In the current study, videos appeared to be recorded using a variety of technologies, including digital camcorders, computer webcams, and cellphones supported by a tripod. Universities may also provide media resources for students without access to appropriate technology.

Video demonstration of error contrasted with correct administration is another potentially helpful way cognitive assessment course instructors could maximize use of video technology and incorporate the results of this study. The WJ-III COG training

materials already include a sample video demonstrating correct administration. In this study's assessment course, portions were shown in class and students were asked to review the video on reserve in the library. This process could be improved as a teaching tool by contrasting correct and incorrect administration, guided by knowledge of the most frequently committed errors. An actor could be recorded administering problematic portions of the test incorrectly. These clips could then be shown in class during course instruction. Students would have the opportunity to identify and discuss the committed error, perhaps including discussion of how the error violates standardization and its potential impact on examinee scores. Another video clip could then be shown of correct administration. Such a presentation could also be used as training materials for teaching assistants to ensure that they are vigilant for commonly committed errors. If feedback based on videotaped administration is to be used as an instructional tool, supervisors need a high level of competence in the test to ensure that feedback is comprehensive and accurate.

Importance of Administering to Varied Examinees. Students will benefit from the opportunity to administer the WJ-III COG to volunteers of varying ages, and course instructors may consider requiring videotaped administration with younger children. A moderate negative correlation was observed between total number of errors and examinee age, primarily due to differences in error rates on certain corrective feedback procedures, suggesting that students have less opportunity to learn from making errors when testing older volunteers who require less correction. Moreover, on some WJ-III COG subtests administration is quite different at the lowest start points (e.g., Visual Matching, Concept Formation). Students who only practice administering the test to adults or adolescents of

average cognitive ability may find themselves committing more errors when in the field they invariably test younger children or individuals with intellectual disabilities. In addition, these populations are more likely to engage in behaviors requiring management and redirection, further increasing the risk of violating standardization if the examiner is ill-prepared. Students completing the cognitive assessment course with no experience in more challenging testing situations may find themselves struggling working with referred children in an environment with less supervision than the cognitive assessment course.

Limitations

This study has a number of limitations. Despite efforts to recruit a more diverse sample, the WJ-III COG administrations analyzed in this study were conducted by students in one school psychology program who all received similar instruction. Moreover, the sample size was small (15 students and 34 videos). The primary effect of this limitation is that results may not be generalizable to students attending other universities. Studies of Wechsler protocols from various university programs suggest that students generally commit similar errors regardless of course instruction, but it is unknown if this finding applies to errors committed on videos. Small sample size also prevented analysis of one of the original aims of this research study, to determine if error rates decreased across the three test administrations. A significant methodological issue noted by Platt et al. (2007) would have also affected investigation of this question, namely that delivery of corrective feedback to students was not uniform. They argue that the only reason for examining test-by-test improvement is if students received comparable feedback before administering each subsequent test. However, several different individuals (the course instructor as well as advanced doctoral student

supervisors) provided feedback based on videos, and these individuals may have varied in their provision of accurate and comprehensive feedback. Furthermore, while the intention was for students to receive feedback before administering the next test, there was no assurance that this occurred in every case.

Another significant limitation is the inclusion of only the first seven subtests of the Standard Battery. Examiners who use the WJ-III COG battery likely give these subtests most frequently as they allow for computation of the General Intellectual Ability score. However, the WJ-III COG is designed for selective testing and includes an additional 13 subtests. Practitioners using cross-battery assessment models in particular are likely to administer additional subtests as needed based on CHC factors (Flanagan, Ortiz, & Alfonso, 2013). The course instructor involved with this study previously attempted to have her graduate students practice almost all the subtests with volunteers, but this requirement created excessively long administration times that were burdensome or aversive for some volunteers.

Low interrater agreement was also problematic. IRA in previous research, when reported, has been high, with simple agreement figures exceeding 90% on studies of Wechsler scales (Alfonso et al., 1998; Loe et al., 2007; Mrazik et al., 2010) and a Cohen's kappa of 0.77 on Loe's (2014) study of the RIAS. Ramos et al. (2009) did not calculate IRA on their study of the WJ-III COG. While an additional check was conducted by a third rater suggesting that my observations were more accurate than those from the second rater, the reliability of the results may still be called into question. Further practice with the second rater before he began scoring videos may have helped prevent this problem, as could a more comprehensive study checklist.

This study, while providing a more complete analysis of examiner error, may still underestimate error in certain ways. Recommended administration procedures that were not easily observed and quantifiable were not analyzed; for example, the publisher's *WJ-III COG Examiner Training Checklist* (2001) includes items such as "Keeps the Test Record behind the Test Book and outside the subject's view," "Communicates to the subject that the session is enjoyable," "Moves smoothly from one test to another," and "Encourages effort and praises the subject for putting forth his or her best effort." While these recommendations are not necessarily errors, they are part of overall test administration proficiency and could certainly have an impact on examinee performance.

Implications for Further Research

As the first study to analyze videotaped administrations, the most needed area for future research is extension of similar methodology to other tests. The WJ-III COG's popularity has grown rapidly in recent years, but the Wechsler scales remain the most frequently used by examiners (Barak, 2008; Dietz, 2012; Orlovsky et al., 2005). The WJ-III COG was chosen for the current study partly for convenience reasons: The course instructor required three WJ-III COG videos but only one WISC-IV video. Course instructors need more accurate information about commonly committed errors on the three Wechsler scales, as well as other frequently utilized measures including the SB5, the DAS-II, and the KABC-2. Importantly, the Woodcock-Johnson battery will be updated to the WJ-IV in late summer 2014. Information available at the time of writing indicates that this revision represents a major change in the test's structure (Houghton Mifflin Harcourt, 2014). Four subtests have been added to the Standard Battery, while Visual-Auditory Learning has been moved to the Extended Battery, among many other

changes. The extent to which administration procedures have changed, and the application of my findings to this revision, are unknown. Further research will be needed when the WJ-IV is available.

Few recent studies of examiner error have included practitioners, perhaps due to recruitment difficulties and the increased ethical concerns associated with identifying errors on test administrations for clinically referred examinees rather than practice volunteers. McDermott, Watkins, and Rhoad's (2014) HLM analysis of a large school-age sample's derived WISC-IV scores suggests that examiner factors, perhaps including systematic administration error, play a greater role in variation than examinees' individual differences. The extant literature focusing on practitioners, based on actual records (Slate et al., 1992; Slate et al., 1993) and contrived protocols (Ryan & Schnakenberg-Ott, 2003; Brazelton et al., 2003), suggests that error rates are fairly similar for students and practitioners on the Wechsler scales. My results may further challenge the common practice of practitioners learning new revisions of tests, or even tests completely new to the practitioner, without formal training (Chattin & Bracken, 1989). Certainly, peer review during the test learning process is needed if practitioners are to competently administer new tests. To understand the depth of the problem of examiner error, additional research with practitioners is needed.

Given the importance of observing student administrations, updated and more specific data are needed on the use of videos and live observations in graduate courses. While previous researchers have provided some information about this topic (Barak, 2008; Cody & Prieto, 2000), disaggregated data based on surveys and/or course syllabi across professional psychology programs concerning the use of these methods may help

further elucidate the significance and prevalence of examiner error as a systemic problem in assessment.

Other areas for future research remedy limitations of this study. Broadening participation to multiple programs and moving beyond convenience sampling to more robust procedures would allow for both improved generalizability of results and systematic examination of differences in cognitive assessment course instruction. A larger sample is also needed to explore the effect of repeated practice on error rates—however, Platt's et al.'s (2007) cautions should be considered during study design in light of my results, specifically assurance that student feedback is given consistently prior to each successive administration and the need to include examinee performance as a covariate. The latter consideration is especially important given the association I observed between examinee age and opportunity to commit error.

Conclusion

This study provides students, practitioners, and cognitive assessment course instructors with a comprehensive analysis of the most common errors committed by graduate students practicing administration of the WJ-III COG. Clinicians use this test, like other standardized intelligence tests, to make significant decisions regarding treatments and services appropriate for the examinee. Given the complexities inherent in psychological assessment, it is imperative that examiner error not be an additional source of uncertainty. Best practice in test administration is nothing less than strictly following standardized procedures, and best practice in graduate preparation requires students to be competent in this area. Further work is needed to ensure that psychologists, and their

educators, have the preparation they need to conduct defensible assessments that accurately assess client needs.

References

- Alfonso, V. C., Johnson, A., Patinella, L., & Rader, D. E. (1998). Common WISC-III examiner errors: Evidence from graduate students in training. *Psychology in the Schools, 35*, 119-125.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- American Psychological Association [APA] Commission on Accreditation. (2013). *Guidelines and principles for accreditation of programs in professional psychology*. Washington, DC: Author. Retrieved from <http://www.apa.org/ed/accreditation/about/policies/guiding-principles.pdf>
- Association of Psychology Postdoctoral and Internship Centers [APPIC]. (2011). *2011 APPIC Match: Survey of internship applicants*. Retrieved from <http://www.appic.org/Match/MatchStatistics/ApplicantSurvey2011Part1.aspx>
- Atkins v. Virginia*, 536 U.S. 304 (2002).
- Babad, E., Mann, M., & Mar-Haylm.(1975). Bias in scoring the WISC subtests. *Journal of Consulting and Clinical Psychology, 43*, 268. doi:10.1037/h0076368
- Barak, A. L. (2008). *Cognitive assessment training in graduate programs in school psychology* (Unpublished doctoral dissertation). St. John's University, Queens, NY.
- Belk, M. S., LoBello, S. G., Ray, G. E., & Zachar, P. (2002). WISC-III administration, clerical, and scoring errors made by student examiners. *Journal of Psychoeducational Assessment, 20*(3), 290-300. doi:10.1177/073428290202000305

- Braden, J. P., & Alfonso, V. C. (2003). The WJ-III in cognitive assessment courses. In F. A. Schrank & D. P. Flanagan (Eds.), *WJ-III Clinical Use and Interpretation: Scientist-Practitioner Perspectives* (pp. 377-396). New York: Academic Press.
- Bramlett, R. K., Murphy, J. J., Johnson, J., Wallingsford, L., & Hall, J. D. (2002). Contemporary practices in school psychology: A national survey of roles and referral problems. *Psychology in the Schools, 39*, 327-335. doi:10.1002/pits.10022
- Brazelton, E. W., 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. *The Professional Educator, 25*(2), 1-8.
- Bub, K. L., Buckhalt, J. A., & El-Sheikh, M. (2011). Children's sleep and cognitive performance: A cross-domain analysis of change over time. *Developmental Psychology, 47*, 1504-1514. doi:10.1037/a0025535
- Cashel, M. L. (2002). Child and adolescent psychological assessment: Current clinical practices and the impact of managed care. *Professional Psychology Research and Practice, 33*, 446-453. doi:10.1037/0735-7028.33.5.446
- Castillo, J. M., Curtis, M. J., & Gelley, C. (2012). School psychology 2010: School psychologists' professional practices and implications for the field. *Communiqué, 40*(8), 4-6.
- Chappell, B. (2014, May 27). Florida's IQ limit for death penalty isn't constitutional, Supreme Court says. *NPR*. Retrieved from <http://www.npr.org/blogs/thetwo-way/2014/05/27/316315861/florida-s-iq-limit-for-death-penalty-isnt-constitutional-supreme-court-says>

- Chattin, S. H., & Bracken, B. A. (1989). School psychologists' evaluation of the K-ABC, McCarthy Scales, Stanford-Binet IV, and WISC-R. *Journal of Psychoeducational Assessment*, 7, 112-130. doi:10.1177/073428298900700202
- Castillo, J. M., Curtis, M. J., & Gelley, C. (2012). School psychologists' professional practices and implications for the field. *Communiqué*, 40(8), 4-6.
- Clemence, A. J., & Handler, L. (2001). Psychological assessment on internship: A survey of training directors and their expectations for students. *Journal of Personality Assessment*, 76, 18-47. doi:10.1207/S15327752JPA7601_2
- Cody, M. S. & Prieto, L. R. (2000). Teaching intelligence testing in APA-accredited programs: A national survey. *Teaching of Psychology*, 27, 190-194.
- Data Accountability Center. (2011). Table B1-3. Number of students ages 6 through 21 served under IDEA, Part B, by disability and state: Fall 2011. In *Individuals with Disabilities Education Act Data*. Retrieved from <https://www.ideadata.org/TABLES35TH/B1-3.pdf>
- Davidson, J. E. & Kemp, I. A. (2011). Contemporary models of intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge Handbook of Intelligence* (pp. 58-84). New York: Cambridge University Press.
- Dietz, V. A. (2012). *A survey of New Jersey school psychologists regarding the measures they use to assess students* (Unpublished doctoral dissertation). Rutgers University, Rutgers, NJ.
- Dombrowski, S. C. (2013). Investigating the structure of the WJ-III Cognitive at school age. *School Psychology Quarterly*, 28, 154-69. doi:10.1037/spq0000010

- Egan, P. J., McCabe, P., Semenchuk, D., & Butler, J. (2003) Using portfolios to teach test-scoring skills: A preliminary investigation. *Teaching of Psychology, 30*, 233-235.
- Flanagan, D. P., Fiorello, C. A., & Ortiz, S. O. (2010). Enhancing practice through application of Cattell–Horn–Carroll theory and research: A “third method” approach to specific learning disability identification. *Psychology in the Schools, 47*, 739-760. doi: 10.1002/pits.20501
- Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2008). Response to intervention (RTI) and cognitive testing approaches provide different but complementary data sources that inform SLD identification. *Communiqué, 36*(5), 16-17.
- Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). Hoboken, NJ: Wiley.
- Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Dynda, A. M. (2006). Integration of response to intervention and norm-referenced tests in learning disability identification: Learning from the Tower of Babel. *Psychology in the Schools, 43*(7), 807-825. doi:10.1002/pits.20190
- Floyd, R. G., Keith, T. Z., Taub, G. E., & McGrew, K. S. (2007). Cattell-Horn-Carroll cognitive abilities and their effects on reading decoding skills: *g* has indirect effects, more specific abilities have direct effects. *School Psychology Quarterly, 22*, 200-233. doi:10.1037/1045-3830.22.2.200
- Franklin, M. R., Stillman, P. L., Burpeau, M. Y., & Sabers, D. L. (1982). Examiner error in intelligence testing: Are you a source? *Psychology in the Schools, 19*, 563-569.

- Fuchs, D., Compton, D. L., Fuchs, L. S., Bryant, V. J., Hamlett, C. L., & Lambert, W. (2012). First-grade cognitive abilities as long-term predictors of reading comprehension and disability status. *Journal of Learning Disabilities, 45*, 217-231. doi:10.1177/0022219412442154
- Glutting, J.J., Watkins, M. W., & Youngstrom, E. A. (2003). Factored and cross-battery ability assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (pp. 343-374). New York: Guilford.
- Goldstein, S. (2013). The science of intelligence testing: Commentary on the evolving nature of interpretations of the Wechsler scales. *Journal of Psychoeducational Assessment, 31*, 132-137. doi:10.1177/0734282913478033
- Gresham, F. M., Restori, A. F., & Cook, C. R. (2008). To test or not to test: Issues pertaining to response to intervention and cognitive testing. *Communiqué, 37*, 5-7.
- Gresham, F. M., & Vellutino, F. R. (2010). What is the role of intelligence in the identification of specific learning disabilities? Issues and clarifications. *Learning Disabilities Research & Practice, 25*, 194-206. doi:10.1111/j.1540-5826.2010.00317.x
- Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, N.J.: Wiley.
- Gurley, J. R. (2008). *An examination of scoring accuracy on intelligence and achievement measures* (Unpublished doctoral dissertation). Sam Houston State University, Huntsville, TX.

Hale, J., Alfonso, V., Berninger, V., Bracken, B., Christo, C., Clark, E., ... Yalof, J.

(2010). Critical issues in response-to-intervention, comprehensive evaluation, and specific learning disabilities identification and intervention: An expert white paper consensus. *Learning Disability Quarterly*, *33*, 223-236.

doi:10.1177/073194871003300310

Hall v. Florida, 572 U.S. ____ (2014).

Hanna, G. S., Bradley, F. O., & Holen, M. C. (1982). Estimating major sources of measurement error in individual intelligences scales: Taking our heads out of the sand. *Journal of School Psychology*, *19*, 370-376. doi:10.1016/0022-4405(81)90031-5

Hinnant, J. B., El-Sheikh, M., Keiley, M., & Buckhalt, J. A. (2013). Marital conflict, allostatic load, and the development of children's fluid cognitive performance. *Child Development*, *84*, 2003-2014. doi:10.1111/cdev.12103

Houghton Mifflin Harcourt. (2014). *Introducing the Woodcock-Johnson IV*. Retrieved from http://www.riversidepublishing.com/products/wj-iv/pdf/96730_Woodcock_Johnson_Newsletter_HR.pdf

Hunnicut, L. C., Slate, J. R., Gamble, C., & Wheeler, M. S. (1990). Examiner errors on the Kaufman Assessment Battery for Children: A preliminary investigation. *Journal of School Psychology*, *28*, 271-278. doi:10.1016/0022-4405(90)90017-2

Jacobs, K. E., Szer, D., & Roodenburg, J. (2012). The moderating effect of personality on the accuracy of self-estimates of intelligence. *Personality and Individual Differences*, *52*, 744-749. doi:10.1016/j.paid.2011.12.040

- Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2007). Response to intervention at school: The science and practice of assessment and intervention. In *Handbook of Response to Intervention* (pp. 3-9). New York: Springer.
- Kamphaus, R. W., Petoskey, M. D., & Rowe, E. W. (2000). Current trends in psychological testing of children. *Professional Psychology Research and Practice*, *31*, 155-164. doi:10.1037/0735-7028.31.2.155
- Kavale, K. A., & Spaulding, L. S. (2008). Is response to intervention good policy for specific learning disability? *Learning Disabilities Research & Practice*, *23*, 169-179. doi:10.1111/j.1540-5826.2008.00274.x
- Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence*, *36*, 502-525.
doi:10.1016/j.intell.2007.11.001
- Klassen, R. M., Neufeld, P., & Munro, F. (2005). When IQ is irrelevant to the definition of learning disabilities: Australian school psychologists' beliefs and practice. *School Psychology International*, *26*, 297-316.
doi:10.1177/0143034305055975
- Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., ... Benton, S. A. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology*, *60*, 725-739.
doi:10.1002/jclp.20010

- Loe, S. A. (2014). Examiner errors on the Reynolds Intellectual Assessment Scales committed by graduate student examiners. *Psychology in the Schools, 51*, 97–106. doi:10.1002/pits.21738
- Loe, S. A., Kadlubek, R. M., & Marks, W. J. (2007). Administration and scoring errors on the WISC-IV among graduate student examiners. *Journal of Psychoeducational Assessment, 25*, 237-247. doi:10.1177/0734282906296505
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1-10. doi:10.1016/j.intell.2008.08.004
- McGrew, K. S., & Wendling, B. J. (2010). Cattell–Horn–Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*, 651-675. doi:10.1002/pits.20497
- McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014) Whose IQ is it?: Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment, 26*, 207-214.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*, 276-282.
- McQuade, J. D., Tomb, M., Hoza, B., Waschbusch, D. A., Hurt, E. A., & Vaughn, A. J. (2011). Cognitive deficits and positively biased self-perceptions in children with ADHD. *Journal of Abnormal Child Psychology, 39*, 307-319. doi:10.1007/s10802-010-9453-7

- McQueen, W., Meschino, R., Pike, P., & Poelstra, P. (1994). Improving graduate student performance in cognitive assessment: The saga continues. *Professional Psychology: Research and Practice, 25*, 283-287. doi:10.1037/0735-7028.25.3.283
- Miller, C. K., Chansky, N. M., & Gredler, G. R. (1970). Rater agreement on WISC protocols. *Psychology in the Schools, 7*, 190-193.
- Mpofu, E., & Oakland, T. (Eds). (2010). *Rehabilitation and health assessment: Applying ICF guidelines*. New York, NY: Springer.
- Mrazik, M., Janzen, T. M., Dombrowski, S. C., Barford, S. W., & Krawchuk, L. L. (2012). Administration and scoring errors of graduate students learning the WISC-IV: Issues and controversies. *Canadian Journal of School Psychology, 27*, 279-290. doi:10.1177/0829573512454106
- National Association of School Psychologists [NASP]. (2010a). *Standards for graduate preparation of school psychologists*. Bethesda, MD: Author. Retrieved from http://www.nasponline.org/standards/2010standards/1_Graduate_Preparation.pdf
- National Association of School Psychologists [NASP]. (2010b). *Principles for professional ethics*. Bethesda, MD: Author. Retrieved from http://www.nasponline.org/standards/2010standards/1_%20Ethical%20Principles.pdf
- Norcross, J. C., & Karpiak, C. P. (2012). Clinical psychologists in the 2010s: 50 years of the APA Division of Clinical Psychology. *Clinical Psychology: Science and Practice, 19*, 1-12. doi:10.1111/j.1468-2850.2012.01269.x
- Orlovsky, K. L., Alfonso, V. C., & Kestenberg, L. B. (2005, August). *Current university-based assessment center testing trends*. Paper presented at the American

Psychological Association's 2005 Annual Convention, Washington, D.C. PDF
retrieved from

http://www.fordham.edu/images/academics/education/hagin_consultation_center/kristin_apa_postervca%5B1%5D.pdf

Pickren, W. E., & Rutherford, A. (2010). *A history of modern psychology in context*. Hoboken, N.J.: Wiley.

Platt, T. L., Zachar, P., Ray, G. E., Lobello, S. G., & Underhill, A. T. (2007). Does Wechsler Intelligence Scale administration and scoring proficiency improve during assessment training? *Psychological Reports, 100*, 547-555.
doi:10.2466/PRO.100.2.547-555

Plumb, G. R., & Charles, D. C. (1955). Scoring difficulty of Wechsler comprehension responses. *Journal of Educational Psychology, 46*, 179-183.
doi:10.1037/h0046974

Ramos, E., Alfonso, V. C., & Schermerhorn, S. M. (2009). Graduate students' administration and scoring errors on the Woodcock-Johnson III Tests of Cognitive Abilities. *Psychology in the Schools, 46*, 650-657.
doi:10.1002/pits.20405

Reason, J. T. (1990). *Human error*. New York: Cambridge University Press.

Reason, J. T. (1997). *Managing the risks or organizational accidents*. Burlington, VT: Ashgate.

Reason, J. T. (2000). Human error: models and management. *British Medical Journal, 320*, 768-770. doi:10.1136/bmj.320.7237.768

- Reason, J. T. (2008). *The human contribution: Unsafe acts, accidents, and heroic recoveries*. Burlington, VT: Ashgate
- Reschly, D. J. (2000). The present and future status of school psychology in the United States. *School Psychology Review, 29*, 507-522.
- Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology, 51*, 535-555.
doi:10.1016/j.jsp.2013.02.003
- Ryan, J. J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R. *Journal of Consulting and Clinical Psychology, 51*, 149-150. doi:10.1037/0022-006X.51.1.149
- Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale - Third Edition. *Assessment, 10*, 151-159.
doi:10.1177/1073191103010002006
- Sattler, J. M. (2008). *Assessment of children: Cognitive applications*. San Diego. Jerome M. Sattler, Publisher, Inc.
- Schwinn, S. D. (2014). How can states measure mental retardation when imposing the death penalty? *Constitutional Law Prof Blog* [Web log post]. Retrieved from <http://lawprofessors.typepad.com/conlaw/2014/02/how-can-states-measure-mental-retardation-when-imposing-the-death-penalty.html>
- Slate, J. R., & Hunnicutt, L. C. (1988). Examiner errors on the Wechsler scales. *Journal of Psychoeducational Assessment, 6*, 280-288. doi:10.1177/073428298800600311

- Slate, J. R., & Jones, C. H. (1989). Can teaching of the WISC-R be improved?: Quasi-experimental exploration. *Professional Psychology: Research and Practice, 20*, 408-410. doi:10.1037/0735-7028.20.6.408
- Slate, J. R., & Jones, C. H. (1990a). Student error in administering the WISC-R: Identifying problem areas. *Measurement & Evaluation in Counseling & Development, 23*, 137-141.
- Slate, J. R., & Jones, C. H. (1990b). Examiner errors on the WAIS-R: A source of concern. *Journal of Psychology, 124*, 343-345.
doi:10.1080/00223980.1990.10543229
- 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. doi:10.1016/0022-4405(92)90021-V
- Slate, J. R., Jones, C. H., & Murray, R. A. (1991). Teaching administration and scoring of the Wechsler Adult Intelligence Scale-Revised: An empirical evaluation of practice administrations. *Professional Psychology: Research and Practice, 22*, 375-379.
doi:10.1037/0735-7028.22.5.375
- 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*, 156-162.
- Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2001). The current status of psychological assessment training in graduate and professional schools. *Journal of Personality Assessment, 77*, 398-407. doi:10.1207/S15327752JPA7703_02

- Sternberg, R. J., & Grigorenko, E. L. (2002). Difference scores in the identification of children with learning disabilities: It's time to use a different method. *Journal of School Psychology, 40*, 65-83. doi:10.1016/S0022-4405(01)00094-2
- Stinnett, T. A., Havey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12*, 331-350. doi:10.1177/073428299401200403
- Tharinger, D., Pryzwansky, W., & Miller, J. (2008). School psychology: A specialty of professional psychology with distinct competencies and complexities. *Professional Psychology: Research and Practice, 39*(5), 529-536. doi:10.1037/0735-7028.39.5.529
- Turner, S. M., DeMers, S. T., Fox, H. R., & Reed, G. M. (2001). APA's guidelines for test user qualifications. *American Psychologist, 56*, 1099-1113. doi:10.1037/0003-066X.56.12.1099
- Van Noord, R. G., & Prevatt, F. F. (2002). Rater agreement on IQ and achievement tests: Effect on evaluations of learning disabilities. *Journal of School Psychology, 40*, 167-176. doi:10.1016/S0022-4405(02)00091-2
- Vellutino, F. R., Scanlon, D. M., Zhang, H., & Schatschneider, C. (2008). Using response to kindergarten and first grade intervention to identify children at-risk for long-term reading difficulties. *Reading and Writing, 21*(4), 437-480. doi:10.1007/s11145-007-9098-2
- Viera, A., J., & Garrett J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*, 360-363.

- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – Fourth Edition: Technical and interpretive manual*. San Antonio, TX: Psychological Corporation.
- Weiner, I. B. (2012). Education and training in clinical psychology: Correcting some mistaken beliefs. *Clinical Psychology: Science and Practice*, 19, 13-16.
doi:10.1111/j.1468-2850.2012.01270.x
- Whitten, J., Slate, J. R., Jones, C. H., Shine, A. E., & Raggio, D. (1994). Examiner errors in administering and scoring the WPPSI-R. *Journal of Psychoeducational Assessment*, 12(1), 49-54. doi:10.1177/073428299401200105
- Willis J. O., Dumont, R., & Kaufman, A. S. (2010). Factor-analytic models of intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge Handbook of Intelligence* (pp. 39-57). New York: Cambridge University Press.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Achievement*. Itasca, IL: Riverside Publishing.
- Woodcock, R.C., McGrew, K.S., & Mather, N. (2007). *Woodcock-Johnson III Tests of Cognitive Abilities, Normative Update*. Itasca, IL: Riverside Publishing.
- Woods, D. D., Dekker, S., Cook, R. Johannesen, L., & Sarter, N. (2010). *Behind human error* (2nd ed.). Burlington, VT: Ashgate.

Participant #: _____
Administration #: _____

Appendix A

WJ-III COG Study Checklist

General	# of Errors (tally)	Video time for each error	<i>If applicable</i>	
			Likely to inflate score?	Likely to deflate score?
Introduces test in some way (as instructed on p. iii, verbatim not required)			<input type="checkbox"/>	<input type="checkbox"/>
Develops a seating arrangement in which the subject can only see the subject's pages but the examiner can see both sides of the Test Book			<input type="checkbox"/>	<input type="checkbox"/>
When testing backwards to obtain the basal, starts with the first item on the preceding page and presents all items on the page if stimuli are visible to the subject			<input type="checkbox"/>	<input type="checkbox"/>
Administers all items on a page when stimuli are visible to the subject rather than stopping in the middle of a page when a ceiling is reached			<input type="checkbox"/>	<input type="checkbox"/>

Test 1: Verbal Comprehension

Follows correct procedure if subject answers sample items incorrectly			<input type="checkbox"/>	<input type="checkbox"/>
Correctly applies ceiling rule			<input type="checkbox"/>	<input type="checkbox"/>
Correctly applies basal rule			<input type="checkbox"/>	<input type="checkbox"/>
Accepts responses that differ in tense or number as correct			<input type="checkbox"/>	<input type="checkbox"/>

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject's side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Participant #: _____
Administration #: _____

Counts responses that are different parts of speech as incorrect			<input type="checkbox"/>	<input type="checkbox"/>
Accepts correct responses in languages other than English			<input type="checkbox"/>	<input type="checkbox"/>
Requests one-word responses when examinee provides two or more, unless otherwise noted in the Test Book			<input type="checkbox"/>	<input type="checkbox"/>
Antonyms: Asks for another answer if subject gives same stimulus word preceded by "non-" or "un-", unless otherwise noted in Test Book			<input type="checkbox"/>	<input type="checkbox"/>
Reads analogies with proper phrasing on Verbal Analogies			<input type="checkbox"/>	<input type="checkbox"/>
Other:				

Test 2: Visual-Auditory Learning

Begins with Introduction 1 for all subjects			<input type="checkbox"/>	<input type="checkbox"/>
Administers Test Story 1 to all subjects			<input type="checkbox"/>	<input type="checkbox"/>
Discontinues testing when a cutoff criterion is met			<input type="checkbox"/>	<input type="checkbox"/>
Makes sure subject verbalizes each symbol when introduced			<input type="checkbox"/>	<input type="checkbox"/>
Does not allow subject to practice or review symbols; turns page immediately after introducing symbols			<input type="checkbox"/>	<input type="checkbox"/>

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject's side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Participant #: _____
Administration #: _____

Allows 5 seconds (± 1 second) for subject to respond before immediately pointing to the symbol and providing the word			<input type="checkbox"/>	<input type="checkbox"/>
Moves immediately to next symbol after providing subject with a word			<input type="checkbox"/>	<input type="checkbox"/>
On the protocol, circles each word that is missed or that is told to the subject			<input type="checkbox"/>	<input type="checkbox"/>
Does not accept synonyms as correct responses			<input type="checkbox"/>	<input type="checkbox"/>
Counts extra words as errors			<input type="checkbox"/>	<input type="checkbox"/>
Queries skipped symbols by pointing to the symbol and saying, "What is this?"			<input type="checkbox"/>	<input type="checkbox"/>
Uses hand or paper to uncover one line at a time if subject requires this accommodation			<input type="checkbox"/>	<input type="checkbox"/>
Other:				

Test 3: Spatial Relations

Begins with Introduction and sample items for all subjects			<input type="checkbox"/>	<input type="checkbox"/>
Gives corrective feedback on Sample Items A through D as instructed			<input type="checkbox"/>	<input type="checkbox"/>
Assigns 1 point for each piece identified correctly			<input type="checkbox"/>	<input type="checkbox"/>
Discontinues testing when a cutoff criterion is met			<input type="checkbox"/>	<input type="checkbox"/>

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject's side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Participant #: _____
Administration #: _____

Encourages subject to identify pieces by naming letters rather than pointing if they know the alphabet			<input type="checkbox"/>	<input type="checkbox"/>
<u>Through Item 22</u> , asks “And what else?” if subject names two pieces when three are needed			<input type="checkbox"/>	<input type="checkbox"/>
Uses hand or paper to uncover one line at a time if subject requires this accommodation			<input type="checkbox"/>	<input type="checkbox"/>
Other:				

Test 4: Sound Blending

Begins with Sample Item A for all subjects			<input type="checkbox"/>	<input type="checkbox"/>
Presents Sample Item A orally and Sample Item B and all test items using the audio recording			<input type="checkbox"/>	<input type="checkbox"/>
Presents additional sample items if subject does not initially understand task				
Presses the pause control button on the audio equipment if a subject needs additional time (<i>count one error per item when audio resumes while subject is still responding</i>)			<input type="checkbox"/>	<input type="checkbox"/>
Looks away from the subject when an audio-recorded test item is being presented, and then looks back as soon as the prompt is heard			<input type="checkbox"/>	<input type="checkbox"/>

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject's side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Participant #: _____
Administration #: _____

Accepts only words pronounced smoothly, not phoneme by phoneme, as correct			<input type="checkbox"/>	<input type="checkbox"/>
Provides one reminder, <i>but no more</i> , about saying word smoothly during test			<input type="checkbox"/>	<input type="checkbox"/>
Does not repeat any items			<input type="checkbox"/>	<input type="checkbox"/>
Presents Items 1 through 16 orally if subject is not responsive to audio recording			<input type="checkbox"/>	<input type="checkbox"/>
Other:				

Test 5: Concept Formation

Begins with Introduction 1 (Preschool to Grade 1) or Introduction 2 (Grade 2 and above)			<input type="checkbox"/>	<input type="checkbox"/>
Discontinues testing when a cutoff criterion is met			<input type="checkbox"/>	<input type="checkbox"/>
Queries responses as appropriate			<input type="checkbox"/>	<input type="checkbox"/>
Accepts correct synonyms (e.g., “small” for <i>little</i> , “circle” for <i>round</i>)			<input type="checkbox"/>	<input type="checkbox"/>
Uses subject’s synonyms when providing corrective feedback (<i>one error per incorrectly-provided feedback</i>)			<input type="checkbox"/>	<input type="checkbox"/>

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject’s side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Participant #: _____
Administration #: _____

Acknowledges correct responses through Item 35 by saying "Right," "Good," "That's correct," or by nodding head			<input type="checkbox"/>	<input type="checkbox"/>
Varies manner of acknowledging correct responses (<i>1 error if acknowledgement is the same word during entire subtest</i>)			<input type="checkbox"/>	<input type="checkbox"/>
Provides corrective feedback on all errors through Item 35			<input type="checkbox"/>	<input type="checkbox"/>
Does not acknowledge correct responses or provide corrective feedback on Items 36-40			<input type="checkbox"/>	<input type="checkbox"/>
Allows only 1 minute each for Items 27-40			<input type="checkbox"/>	<input type="checkbox"/>
Other:				

Test 6: Visual Matching

Administers the appropriate version of the test: VM1 for ages 2-4, VM2 for ages 5+			<input type="checkbox"/>	<input type="checkbox"/>
Adheres to appropriate time limits (+- 3 seconds) for each version: 2 min for VM1, 3min for VM2			<input type="checkbox"/>	<input type="checkbox"/>
Uses a stopwatch or records exact starting and stopping times			<input type="checkbox"/>	<input type="checkbox"/>
After sample items on VM2, holds up protocol so subject cannot study items			<input type="checkbox"/>	<input type="checkbox"/>

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject's side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Participant #: _____
Administration #: _____

Records exact finishing time in minutes and seconds on Test Record if under the time limit			<input type="checkbox"/>	<input type="checkbox"/>
On VM2, prompts immediately if subject tries to stop at bottom of first column			<input type="checkbox"/>	<input type="checkbox"/>
Uses paper or hand to uncover one line at a time <i>only</i> on VM1, if needed			<input type="checkbox"/>	<input type="checkbox"/>
On VM1, turns page immediately after subject responds to last item on that page			<input type="checkbox"/>	<input type="checkbox"/>
Other:				

Test 7: Numbers Reversed

Administers Sample Items A through C and Items 1 through 10 orally			<input type="checkbox"/>	<input type="checkbox"/>
Uses audio recording to administer Sample Item D and all remaining items			<input type="checkbox"/>	<input type="checkbox"/>
Presses the pause control button on the audio equipment if a subject needs additional time (<i>count one error per item when audio resumes while subject is responding</i>)			<input type="checkbox"/>	<input type="checkbox"/>
Looks away from the subject when an audio-recorded test item is being presented, and then looks back as soon as the prompt is heard			<input type="checkbox"/>	<input type="checkbox"/>

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject's side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Participant #: _____

Administration #: _____

Reminds subject to say numbers backwards only when indicated in Test Book			<input type="checkbox"/>	<input type="checkbox"/>
Presents items orally if subject is not responsive to audio recording			<input type="checkbox"/>	<input type="checkbox"/>
Other:				

Common Errors	Tally	Time(s)
Instructions not given verbatim (1 error per blue text segment with at least one word added or omitted)		
Points incorrectly to stimuli on subject's side		
Gives subject inappropriate feedback		
Incorrectly records examinee response on protocol		

Appendix B



INDIANA UNIVERSITY
OFFICE OF RESEARCH ADMINISTRATION

To: JACK ALAN CUMMINGS
EDUCATION

From: IU Human Subjects Office
Office of Research Administration – Indiana University

Date: September 21, 2012

RE: EXEMPTION GRANTED

Protocol Title: A Videotape Analysis of Graduate Students' Administration Errors on the WJ-III COG

Protocol #: 1209009567

Funding Agency/Sponsor: None

IRB: IRB-IUB, IRB00000222

Your study named above was accepted on September 21, 2012 as meeting the criteria of exempt research as described in the Federal regulations at 45 CFR 46.101(b), paragraph(s) (1) . This approval does not replace any departmental or other approvals that may be required.

As the principal investigator (or faculty sponsor in the case of a student protocol) of this study, you assume the following responsibilities:

Amendments: Any proposed changes to the research study must be reported to the IRB prior to implementation. To request approval, please complete an Amendment form and submit it, along with any revised study documents, to irb@iu.edu. Only after approval has been granted by the IRB can these changes be implemented.

Completion: Although a continuing review is not required for an exempt study, you are required to notify the IRB when this project is completed. In some cases, you will receive a request for current project status from our office. If we are unsuccessful at in our attempts to confirm the status of the project, we will consider the project closed. It is your responsibility to inform us of any address changes to ensure our records are kept current.

Per federal regulations, there is no requirement for the use of an informed consent document or study information sheet for exempt research, although one may be used if it is felt to be appropriate for the research being conducted. As such, these documents are returned without an IRB-approval stamp. Please note that if your submission included an informed consent statement or a study information sheet, the IRB requires the investigational team to use these documents.

You should retain a copy of this letter and any associated approved study documents for your records. Please refer to the project title and number in future correspondence with our office. Additional information is available on our website at <http://researchadmin.iu.edu/HumanSubjects/index.html>.

If you have any questions, please contact our office at the below address.

Thank you.

INDIANA UNIVERSITY INSTITUTIONAL REVIEW BOARD (IRB)

EXEMPT RESEARCH CHECKLISTIRB Study Number: 1209009567Principal Investigator: Dr. Jack Cummings / Luke Erichsen (co-PI)Study Title: **A Videotape Analysis of Graduate Students' Administration Errors on the WJ-III COG**Document Date: 9/19/2012

DIRECTIONS: This form is to be neatly typed and submitted to the IRB only when the investigator is contemplating the initiation of a research project which, in the investigator's judgment, is exempt from full IRB review. The IRB will then determine whether the activity is covered by these regulations. *Please type only in the gray boxes. To mark a box as checked, double-click the box, select "checked", and click "OK".*

Research activities are exempt from regulations for the protection of human research subjects when they are considered minimal risk (the probability or magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests (as defined by 45 CFR 46.102(i)) and the ONLY involvement of human subjects falls within one or more of the exempt categories listed below.

The exempt categories outlined below do not apply to research involving prisoners or research involving a test article regulated by the FDA, unless the research meets the criteria for exemption described in 45 CFR 46.101(b)(6) and 21 CFR 56.104(d). Additionally, research involving pregnant women that is conducted at or funded by the VA can not be exempt.

The exempt categories outlined below are based solely on methods of research, and do not take the level of risk into consideration. Although most exempt research requires no further oversight to be conducted ethically, some exempt research raises ethical concerns or requires measures to protect participants. As such, the IRB will not consider any research exempt that does not fulfill ethical principles reflected in the Belmont Report. These basic ethical principles are:

1. Respect for Persons (Autonomy) – individuals should be treated as autonomous agents and persons with diminished autonomy are entitled to protection.
2. Beneficence – Human subjects should not be harmed and the research should maximize possible benefits and minimize possible harms.
3. Justice – the benefits and risks of research must be distributed fairly.

Research that otherwise would be exempt by federal regulations that raises ethical concerns or requires measures to protect subjects may be denied and/or moved to a higher level of review (i.e. expedited or full IRB review).

SECTION I: EXEMPT CATEGORY

Check the appropriate category(ies) that applies to your research project:

<input checked="" type="checkbox"/>	1. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research on regular and special educational instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods. [45CFR46.101(b)(1)]
<input type="checkbox"/>	2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless all of the following are true: <p style="margin-left: 40px;">(i) information obtained is recorded in such a manner that the human subjects can be identified, directly or through identifiers linked to the subjects; and</p> <p style="margin-left: 40px;">(ii) any disclosure of the subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, insurability, or reputation. [45CFR46.101(b)(2)]</p> <p>NOTE: If the research involves children as participants, the research must be limited to educational tests (cognitive, diagnostic, aptitude, achievement) and observation of public behavior when the investigator(s) do</p>

	not participate in the activities being observed. Research involving children that uses survey procedures, interview procedures, or observation of public behavior when the investigator(s) participate in the activities being observed cannot be granted an exemption.
<input type="checkbox"/>	<p>3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under category 2 above, if either:</p> <p>(i) the human subjects are elected or appointed public officials or candidates for public office; or</p> <p>(ii) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter. [45CFR46.101(b)(3)]</p>
If any of the above categories have been selected, answer the following:	
<p>Will you be audio or video recording?</p> <p><input type="checkbox"/> No</p> <p><input checked="" type="checkbox"/> Yes. Explain how it will be assured that the identity of the subjects and/or link to the information obtained or the information recorded about the subjects does not place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, insurability, or reputation:</p> <p>Background: This research analyzes data gathered as part of the normal training process for graduate students in school psychology. First-year students in the school psychology program are required to conduct practice administrations of cognitive tests with children and submit videotapes of these administrations to their course instructor. Results of these practice administrations are not released to parents, are not part of the child's educational record, and are not used in any decision-making capacity.</p> <p>The investigators will be obtaining the videotapes and written test materials from the graduate students who choose to participate and will not have any direct contact with children. <i>The focus of this study is exclusively on the graduate students' behavior, and the investigators will not have access to any of the children's personally identifiable information.</i> Parental consent for participation in the graduate students' required practice administrations for the course is obtained by the graduate students using a form designed by the instructor (<i>attached</i>).</p> <p>Errors in administration are inherently part of the training process and documentation of these errors in a practice administration does not place the graduate student participants at risk in any way.</p> <p>Please see question 5(b) below for more information about maintaining confidentiality of the videos.</p>	
<input type="checkbox"/>	<p>4. Research involving the collection or study of <u>existing</u> data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. [45CFR46.101(b)(4)]</p> <p>To qualify for this exemption, data, documents, records, or specimens must exist at the time the research is proposed and not prospectively collected.</p> <p>Provide a list of all data points (the types of data) that will be collected below or attach a data collection sheet.</p>
<input type="checkbox"/>	<p>5. Research and demonstration projects which are conducted by or subject to the approval of Department or Agency heads, and which are designed to study, evaluate, or otherwise examine:</p> <p>(i) public benefit or service programs;</p> <p>(ii) procedures for obtaining benefits or services under those programs;</p> <p>(iii) possible changes in or alternatives to those programs or procedures; or</p> <p>(iv) possible changes in methods or levels of payment for benefits or services under those programs. [45CFR46.101(b)(5)].</p> <p>The program under study must deliver a public benefit (for example, financial or medical benefits as provided under</p>

	<p>the Social Security Act) or service (for example, social, supportive, or nutrition services as provided under the Older Americans Act).</p> <p>The research or demonstration project must be conducted pursuant to specific federal statutory authority, must have no statutory requirement that an IRB review the project, and must not involve significant physical invasions or intrusions upon the privacy of the subjects.</p> <p>This exemption is for projects conducted by or subject to approval of Federal agencies and requires authorization or concurrence by the funding agency.</p>
<input type="checkbox"/>	<p>6. Taste and food quality evaluation and consumer acceptance studies,</p> <p>(i) if wholesome foods without additives are consumed; or</p> <p>(ii) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural, chemical, or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection Service of the U.S. Department of Agriculture. [45CFR46.101(b)(6) and 21 CFR 56.104(d)]</p>

SECTION II: PERFORMANCE SITE

- Indiana University
 - IUB Campus. Please state school/department/location(s): School of Education, Counseling & Educ. Psychology
 - IUPUI Campus. Please state school/department/location(s): _____
 - Bradford Woods
 - Center for Survey Research
 - Center for Evaluation & Education Policy (CEEP)
 - Indiana Clinical Research Center (ICRC)*
 - Indiana Institute on Disability and Communication
 - IU Simon Cancer Center*
 - Krannert Institute of Cardiology*
 - Kinsey Institute
 - Oral Health Research Institute
 - Other: _____
- Health & Hospital Corporation of Marion County
 - Bell Flower Clinic
 - Midtown Mental Health*
 - Wishard Memorial Hospital*
 - Hospital/ER
 - Non-primary care
 - Wishard Specialty Clinics
 - OB/GYN Clinics
- Indiana University Health (Clarian) Facilities
 - Bloomington Hospital
 - Beltway Centers
 - Methodist Hospital
 - Methodist-Affiliated Centers/Private Practices
 - North Hospital
 - Riley Hospital for Children
 - University Hospital
 - West Hospital
 - Other: _____
- IU Health Clinics. Please list location: _____.
- IU Medical Group Specialty Clinic (IUMG-SC). Please list location: _____.
- Larue Carter Hospital
- Monroe County Community School Corporation. Please list school: _____.
- Regenstrief Institute
- Rehabilitation Hospital of Indiana
- Richard L. Roudebush Veterans Affairs Medical Center*
- Other: _____

* Additional information or submission may be required prior to initiating the study. Please check with the specific performance site for additional information.

Any study using the VA as a performance site, using VA patients, or funded by the VA MUST be submitted to and receive approval from the VA R&D Committee **before any research can be conducted at the VA..

SECTION III: RESEARCH DESCRIPTION
--

NOTE: Study information will be released to the Clinical and Translational Science Institute (CTSI) for the clinical trials listing. To opt out of this listing requirement you will need to get opt-out approval from Dr. Anantha Shekhar, PhD, MD, Director of Indiana CTSI, prior to IRB submission. For additional information or to request opt-out approval, please contact Sam Scahill at (317) 278-6969 or sscahill@iupui.edu.

1. Provide a brief description, in lay terms, of the purpose of the proposed project.

This project will build upon previous research concerning the frequency of error in administration and scoring of standardized cognitive tests, including the Woodcock-Johnson Tests of Cognitive Abilities, 3rd Edition (WJ-III COG). Administering cognitive and achievement tests correctly is essential to ensure results for a child that can serve as a valid comparison with scores obtained from children in the standardization sample. Previous researchers have analyzed errors found on written records of test administrations; the current study will describe and quantify errors made during practice administrations by graduate students that cannot be identified through examination of test records (i.e., errors that are only observable *in vivo* or using videotapes). Training in administration of these tests may be improved by understanding errors commonly made by examiners.

- a. List all methods by which information or data about or from subjects will be obtained. Describe the frequency and duration of the procedures. **NOTE:** Please include all surveys, instruments, survey/interview questions, etc. that will be used for this research.

The participants are already required to submit three videotaped practice administrations of the WJ-III COG to the course instructor, along with associated written test forms. Those who choose to participate in this study will be asked to submit these same videos and forms to the investigators after they have had the opportunity to review feedback from the instructor. The participating students who provide email addresses will all be individually emailed near the end of the semester, reminding them to place their materials in a sealed envelope in the co-PI's mailbox in the School of Education Counseling and Educational Psychology suite. Digital copies of the materials, including scans of the test forms, will be created and stored on Dropbox (as described below in 5b). The original materials will be returned to participants via the suite mailboxes within two weeks.

The videotapes will be coded for administration errors using the attached protocol.

Complete 2-6 below ONLY if you selected Categories 1, 2, 3, 5, or 6 in Section I above.

2. Please state the eligibility (inclusion/exclusion criteria).

Graduate students in the first-year school psychology course on cognitive assessment (EDUC P655) who have no previous experience with the administration of cognitive tests.

3. Will subjects be paid for participation in the study (e.g. monetary, free services, gifts, course credit, including extra credit)?

- No. Proceed to 4.
 Yes. Complete items a. and b. below.

- a. Explain the payment arrangements (e.g. amount and timing of payment and the proposed method of disbursement). **NOTE:** Payments must accrue and not be contingent upon completion of the study. However, a small payment (bonus) for completion of the study may be acceptable if it is found to not be persuasive for the subjects to remain in the study.
- b. Justify the proposed payment arrangements described in section B. (e.g., how this proposed payment arrangement is not considered to be coercive).

4. Provide the process by which individuals will be identified and recruited. **Note:** Please include a copy of all information to be shared with or intended to be seen by potential subjects to inform them of this research and ask for their participation.

All eligible individuals (students in P655 without prior experience in cognitive assessment) will be invited to participate. The co-PI will attend a class to discuss the research project and recruit participants.

The primary participants in this study are the graduate students administering the tests. They are responsible for recruiting children to help them meet a course requirement involving administration of practice tests, and the subjects will not be recruiting additional children for this study beyond what is already required of them.

- a. Explain how it will be ensured that recruitment or selection will not unfairly target a particular population or will target the population that will benefit from the project/research.

All eligible individuals may participate.

5. Explain how it will be ensured that individuals will be treated with respect during interactions/observations with them. For those individuals with diminished autonomy (e.g. children, people with limited ability to make decisions), explain how they will be protected.

There will be limited or no direct contact with the participants after recruitment regarding this study.

To address possible coercion in the use of students, the instructor of the cognitive assessment course will not be an investigator and will be unaware of which students are participating in the study. The co-PI will visit the course to explain the study's purpose and then distribute an informed consent forms to everyone in the class. At the end of the class period, these will be returned, folded, to the co-PI. Students who do not desire to participate may return the informed consent form unsigned. An additional, separate form will also be used to collect email addresses *Students will receive corrective feedback on their performance in the videos from the course instructor regardless of their participation in the study.*

- a. Explain how subject privacy will be protected. For example, if interviewing, where will that be conducted?

The subjects are responsible for conducting the videotaped administrations in private, appropriate settings. The videos will not be viewed for coding purposes in public areas (e.g., computers in computer labs).

Regarding privacy during the consent process, neither the subjects' peers nor the course instructor will know who is participating in the study unless they reveal so themselves.

- b. Explain how subject confidentiality will be protected. For example, what kind of information will be recorded and how will that be protected?

The video files provided by the subjects will be transferred to and stored in a password-protected Dropbox account, which uses encryption technology to protect against unauthorized transmission of data. Original computer media containing the videos, such as CDs or USB flash drives, will be returned to the subjects. ID numbers will be used to avoid recording the graduate students' names on the study protocols used for coding errors (*see attached protocol*). The use of ID numbers will allow the investigators to track an individual participant's progress over the three practice administrations without recording names. The video filenames will also use ID numbers rather than names, and the videos will be promptly erased following conclusion of the study. Written test records associated with the practice administration will also not include personally identifiable information, and the scanned copies of these records will be erased following the conclusion of the study.

- c. Explain how subjects will be fully informed of this research prior to their participation (through the use of a consent form, study information sheet, etc.). **Note:** Please provide a copy of the consent form, study information sheet, etc.

Informed consent forms will be provided to the participants, and the co-PI will also visit the class to explain the study and answer questions. Please refer to the informed consent form for information about study information that will be shared with the participants

6. How will you help to minimize potential risks that individuals may be exposed to while participating in the research? Potential risks may include psychological, social, legal, physical, etc.

No foreseeable risks to participation in this study can be identified.

7. Are you enrolling non-military, non-US research subjects (excluding internet research which may incidentally enroll non-US research subjects)?

No.

Yes. Please describe your familiarity with local customs, culture, and local ethical review requirements: _____

Appendix C

**INDIANA UNIVERSITY****OFFICE OF RESEARCH ADMINISTRATION**

To: JACK ALAN CUMMINGS
EDUCATION

From: IU Human Subjects Office
Office of Research Administration – Indiana University

Date: August 01, 2013

RE: NOTICE OF EXPEDITED APPROVAL - AMENDMENT

Protocol Title: A Videotape Analysis of Graduate Students' Administration Errors on the WJ-III COG

Protocol #: 1209009567

Funding Agency/Sponsor: None

IRB: IRB-IUB, IRB00000222

An amendment to your above-referenced protocol was approved by the Institutional Review Board on July 31, 2013. The protocol meets the requirements for expedited review pursuant to §46.110(b)(2). The changes described in the amendment can now be implemented, unless any departmental or other approvals are required.

If you submitted a revised informed consent document a copy of the approved stamped document is enclosed and must now be used.

You should retain a copy of this letter and any associated approved study documents for your records. All documentation related to this protocol must be maintained in your files for audit purposes for at least three years after closure of the research; however, please note that research studies subject to HIPAA may have different requirements regarding file storage after closure. Additional information is available on our website at <http://researchadmin.iu.edu/HumanSubjects/index.html>. If you have any questions, please contact our office at the below address.

Thank you.

INDIANA UNIVERSITY INSTITUTIONAL REVIEW BOARD (IRB)

STUDY AMENDMENT**Reviewing IRB (please choose one):**

Biomedical: IRB-02 IRB-03 IRB-04 IRB-05
 Behavioral: IRB-01 IUB IRB

IRB STUDY NUMBER: 1209009567AMENDMENT NUMBER: 001

Please type only in the gray boxes. To mark a box as checked, double-click the box, select "checked", and click "OK".

SECTION I: INVESTIGATOR INFORMATION**Principal Investigator:**Name (Last, First, Middle Initial): Cummings, Jack ADepartment: School of Education Phone: 812-856-8327 E-Mail: cummings@indiana.edu**Additional Study Contact:**Name: Luke Erichsen Phone: 812-327-7679 E-Mail: luerichs@indiana.eduProject Title: A Videotape Analysis of Graduate Students' Administration Errors on the WJ-III COG

Sponsor/Funding Agency: _____ Sponsor Number: _____

Sponsor Amendment Number: _____

SECTION II: STUDY INFORMATION

This study is:

- Open to enrollment
 Closed to enrollment

Number of active subjects: 5**SECTION III: AMENDMENT DESCRIPTION**

1. Provide a complete description of the proposed change(s) included in this amendment:

The proposed changes are intended to broaden study participation to other universities. No proposed changes increase risk to the participants. Due to differences in the previously approved study design for IU participants, I am offering payment as an incentive to encourage participation by students from other universities. I am also slightly revising data collection procedures for IU participants to facilitate study completion. In addition to modified data collection procedures, a supplemental informed consent form, and a supplemental recruitment email, a revised study protocol for data analysis is also attached to this amendment. This protocol aids in analysis of the videotapes and makes no substantive changes to the previously approved document.

In addition to recruitment from EDUC-P 655 at Indiana University, participants will be recruited from other school psychology graduate programs. These primarily include other universities with connections to the region or previous researchers pursuing similar lines of inquiry, including

I will contact faculty from these programs to determine who currently teaches the course in cognitive assessment. This instructor will be provided with details about the research and asked to forward study information to students via email. Students who would like to participate will be asked to respond to me by email with a mailing address and will also be encouraged to ask any questions they may have.

These students will be sent via U.S. mail an informed consent form, blank DVD media, a \$10 Amazon.com gift certificate, and a prepaid envelope to return study materials. These include a signed informed consent form, DVD with video file, and copy of the test record with examinee score report. Once these materials are returned, I will send each participant another \$10 Amazon.com gift certificate by email. Participants will be specifically instructed to not include the examinee's full name anywhere on the enclosed returned materials. The test records will be digitized and uploaded along with the video files to Box, a secure IU-sponsored storage service.

Revision to procedures at Indiana University: To maximize study participation at IU, I will again recruit participants in person as previously approved. However, rather than asking participants to directly provide me with their materials, I will coordinate with the course instructor to receive the material packets from all students and as soon as possible copy the needed study data for only those students who consented to participate in the study. As students' names are written on the outside of sealed packets, I will have no access to materials for students who choose not to participate. In this manner the instructor will be continue to be unaware of which students agreed to participate. If this arrangement proves not to be workable, the same data collection procedures as previously approved will be used in which the participants provide the materials directly to me, rather than me directly receiving them from the instructor.

The informed consent form for both populations also now specifies that students may provide an email address, and if they do so, they also consent to receiving two email reminders regarding study participation.

2. State the justification/rationale for this amendment. If risks are being updated, please provide specific justification:
This amendment is needed to modify data collection procedures, including allowing for providing payment.
3. Is the study sponsored?
 - No.
 - Yes. Check the appropriate line below and provide with this amendment, as applicable:
 - A copy of the sponsor's amendment, if the amendment came from the sponsor.
 - A copy of your notice to the sponsor of this change, if you initiated the amendment.
 - A copy of the approved amendment will be sent to the sponsor.
 - None of the above apply. Please explain: _____
4. Do the proposed change(s) described in this amendment alter the risk to benefit assessment?
 - No.
 - Yes. Please describe how the assessment is altered: _____
5. Do the proposed change(s) described in this amendment require changes to the informed consent and/or assent document(s) or process?
 - N/A. Informed consent, written documentation of informed consent, and/or assent has been waived for this study. Skip to Section IV..
 - No. Skip to Section IV.
 - Yes. Answer items A and B below.
 - A. Check the appropriate line below.
 - The new informed consent and/or assent document(s) are in addition to the current one(s).
 - The new informed consent and/or assent document(s) replace the current one(s).
If there are multiple consent and/or documents for this study, please indicate which consent and/or assent document(s) are to be replaced. _____
 - N/A. Changes are being made to the informed consent process only and informed consent document(s) will not change.
 - B. Will enrolled subjects be informed of the change(s) described in this amendment?
 - No. Please explain why not: The changes are only relevant to recruitment to the study. Materials already submitted by subjects will be used as already consented to. The modified study protocol used for analysis of the videos (Examinee Checklist) does not include substantive changes.
 - Yes. Will enrolled subjects be re-consented and/or re-assented?
 - Yes.
 - No. Please explain how enrolled subjects will be notified: _____

SECTION IV: CO-INVESTIGATOR UPDATE

- This submission does NOT include additions or removals to the Investigator List. *Proceed to Section V.*
- This submission includes additions or removals to the Investigator List. The updated Investigator List is attached.

The following investigators are being added to the current Investigator List:

The following investigators are being **removed** from the Investigator List and will no longer be participating in this research:

SECTION V: AMENDMENT SUMMARY

Amendment includes:

- | | |
|---|--|
| <input type="checkbox"/> Assent, dated: _____ | <input checked="" type="checkbox"/> Protocol, dated: _____ |
| <input type="checkbox"/> Number of assent documents: _____ | <input checked="" type="checkbox"/> Recruitment materials (please list and date): <u>Email Recruitment for Non-IU Participants</u> |
| <input type="checkbox"/> Authorization, dated: _____ | <input type="checkbox"/> Request form(s) for vulnerable population(s) (please list and date); _____ |
| <input type="checkbox"/> Number of authorizations: _____ | <input type="checkbox"/> Surveys, questionnaires (please list and date): _____ |
| <input type="checkbox"/> Clinical Investigator's Brochure, dated: _____ | <input type="checkbox"/> Summary Safeguard Statement or HUD Form, dated: _____ |
| <input type="checkbox"/> Expedited Research Checklist, dated: _____ | <input type="checkbox"/> Study Information Sheet |
| <input checked="" type="checkbox"/> Exempt Research Checklist, dated: _____ | <input type="checkbox"/> Other (please list and date): _____ |
| <input type="checkbox"/> HIPAA & Recruitment Checklist, dated: _____ | |
| <input checked="" type="checkbox"/> Informed Consent, dated: _____ | |
| <input type="checkbox"/> Number of consent documents: <u>2</u> | |

Investigator List, dated: _____

NOTE: Only documents that are being changed as a result of the amendment should be attached and checked in items 6 above. Listing document dates are optional and only necessary if required by the investigator or sponsor.

NOTE TO INVESTIGATORS: Study amendments *may not* be instituted until approval from the IRB is given.

Please indicate the type of amendment you are submitting. Please see the Guidelines for Determining an Amendment Type available on the IU Human Subjects Office website for additional information. **Please note that the IRB makes the final determination with regard to whether or not the amendment is acceptable for expedited review or if it requires review at a convened IRB meeting.**

- Minor Amendment.** Change(s) do not significantly affect the safety of subjects and is acceptable for expedited review per 45 CFR 46.110(b)(2)/21 CFR 56.110(b)(2).
- Major Amendment.** Changes potentially involve increased risks or discomforts or decrease potential benefit. The amendment requires review at a convened IRB meeting.

SECTION VI: INVESTIGATOR STATEMENT OF COMPLIANCE

By submitting this form, the Principal Investigator assures that all information provided is accurate. He/she assures that procedures performed under this project will be conducted in strict accordance with federal regulations and Indiana University policies and procedures that govern research involving human subjects. He/she acknowledges that he/she has the resources required to conduct research in a way that will protect the rights and welfare of participants, and that he/she will employ sound study design which minimizes risks to subjects. He/she agrees to submit *any* change to the project (e.g. change in principal investigator, research methodology, subject recruitment procedures, etc.) to the Board in the form of an amendment for IRB approval prior to implementation.

SECTION VII: IRB APPROVAL

This amendment, including documentation noted above, has been reviewed and approved by the Indiana University IRB as meeting the criteria for IRB approval as outlined in 45 CFR 46.111(a). I agree with the investigator's assessment above regarding whether the amendment is a minor or major amendment, unless otherwise noted.

Authorized IRB Signature:  IRB Approval Date: 07.31.13

Printed Name of IRB Member: John R. Baumann

Appendix D

All Errors by Examiner and Administration – Participant 1

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	8y9m	4y9m	9y9m
Verbatim Error	7	16	10
Gave Inappropriate Feedback	1	8	1
Did not record examinee response correctly	9	1	1
Test 1: Verbal Comprehension			
Failed to request 1-word response	1		1
Did not query verbal response		1	
Penalized articulation error		1	
Incorrect ceiling		1	
Mispronounced stimulus word			
Failed to administer sample item			1
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	8	4	3
Failed to point to symbols when correcting	1	2	
Failed to query skipped symbols	1		1
Provided suffix on Story 7 unnecessarily	1		
Test 3: Visual-Spatial Processing			
Violated cutoff rule, did not establish ceiling			
Test 4: Sound Blending			
Did not use audio recording on advanced items			1
Test 5: Concept Formation			
Failed to query responses appropriately	3		
Failed to acknowledge correct responses			3
Test 6: Visual Matching			
Failed to administer VM2 completely after VM1		1	
Failed to hold up book after sample on VM2			1
Let examinee finish item after time limit			1
Test 7: Numbers Reversed			
Presented item backwards		1	
Did not use audio recording on advanced items			1

All Errors by Examiner and Administration – Participant 2

	Administration		
	N/A	N/A	3
<i>Examinee's Reported Age</i>			11y4m
Verbatim Error			14
Did not record examinee response correctly			1
Test 1: Verbal Comprehension			
Failed to request 1-word response			1
Failed to administer sample item			1
Test 2: Visual-Auditory Learning			
Mispronounced stimulus word			2
Test 3: Visual-Spatial Processing			
Failed to give corrective feedback on sample			1
Test 4: Sound Blending			
No Errors			
Test 5: Concept Formation			
Failed to acknowledge correct responses			2
Test 6: Visual Matching			
No Errors			
Test 7: Numbers Reversed			
No Errors			

All Errors by Examiner and Administration – Participant 3

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	15y3m	14y9m	12y8m
Verbatim Error	23	31	31
Gave Inappropriate Feedback	1	2	
Did not record examinee response correctly	8	1	2
Test 1: Verbal Comprehension			
Did not query verbal response	1		
Incorrect ceiling			1
Failed to administer sample item		3	3
Queried clearly incorrect response			1
Gave items in incorrect order	1		
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	4		2
Failed to point to symbols when correcting	10	10	3
Failed to correct an error			1
Provided suffix on Story 7 unnecessarily	1		
Test 3: Visual-Spatial Processing			
Failed to give corrective feedback on samples		2	3
Test 4: Sound Blending			
Failed to allow adequate time for responses		2	
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback	6	1	1
Failed to query responses appropriately			2
Failed to acknowledge correct responses		5	7
Failed to provide corrective feedback on errors		1	3
Provided incorrect feedback on sample		1	
Skipped prompt "next answers have 3 parts"		1	
Failed to ask to repeat sample correctly	2		
Test 6: Visual Matching			
No errors			
Test 7: Numbers Reversed			
Failed to allow adequate time for responses	3		1

All Errors by Examiner and Administration – Participant 4

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	11y0m	17y2m	14y2m
Verbatim Error	9	2	5
Did not record examinee response correctly	2		3
Test 1: Verbal Comprehension			
Did not query verbal response		1	
Did not request 1-word response			1
Failed to reverse at end of page for basal	1	1	1
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	5		
Allowed inappropriate practice on samples		1	
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
Failed to allow adequate time for responses			1
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback	2	1	2
Failed to query responses appropriately			1
Failed to acknowledge correct responses		1	
Test 6: Visual Matching			
Failed to hold up book after sample on VM2		1	
Test 7: Numbers Reversed			
No errors			

All Errors by Examiner and Administration – Participant 5

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	5y11m	6y5m	11y0m
Verbatim Error	30	31	17
Gave Inappropriate Feedback		1	
Did not record examinee response correctly	1	1	
Test 1: Verbal Comprehension			
Did not query verbal response	1	2	
Antonyms: Asks for another if “non/un”		1	
Failed to follow procedure if sample incorrect		1	
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	1	2	
Failed to point to symbols when correcting		9	21
Did not query skipped symbol		1	
Test 3: Visual-Spatial Processing			
Failed to ask examinee to identify by letter	1		
Test 4: Sound Blending			
No errors			
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback	6	1	1
Failed to query responses appropriately		1	
Failed to acknowledge correct responses	3	1	
Corrects a synonym inappropriately	1		1
Test 6: Visual Matching			
No errors			
Test 7: Numbers Reversed			
Gave example for incorrect start point		1	
Presented oral item prompts too fast or slow			1

All Errors by Examiner and Administration – Participant 6

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	15y9m	14y0m	14y4m
Verbatim Error	9	7	2
Gave Inappropriate Feedback		8	1
Did not record examinee response correctly	3	3	2
Test 1: Verbal Comprehension			
Mispronounced stimulus word	2	1	1
Incorrect basal	1		
Accepts incorrect word as correct on sample		1	
Failed to ask for better of two responses		1	
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	1	2	
Failed to point to symbols when correcting	3	2	
Does not allow examinee to verbalize symbols	2		
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
No errors			
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback			2
Failed to query responses appropriately		2	
Failed to acknowledge correct responses	3		
Failed to provide corrective feedback on errors		2	
Test 6: Visual Matching			
Failed to hold up book after sample on VM2	1	1	
Test 7: Numbers Reversed			
Fails to allow adequate time for responses	1	3	3
Cued audio track incorrectly			1
Gave example for incorrect start point	1		
Ceiling error	1		
Presented oral item prompts too fast or slow	6		

All Errors by Examiner and Administration – Participant 7

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	16y6m	9y7m	16y2m
Verbatim error	10	24	8
Gave inappropriate feedback	1	21	6
Did not record examinee response correctly	1	4	
Failed to administer all visible items on page		3	
Test 1: Verbal Comprehension			
Failed to request 1-word response		2	1
Failed to query verbal response	2	2	1
Incorrect ceiling		3	
Failed to reverse at end of page for basal		1	
Antonyms: Asks for another if "non/un"		1	
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors			5
Failed to point to symbols when correcting		3	
Mispronounced stimulus word	1		
Allowed practice on samples		2	
Did not count extra word as error	1		
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
Miscued audio track			1
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback	1		1
Failed to query responses appropriately		2	
Corrected a synonym inappropriately	1		
Failed to provide corrective feedback on errors		4	
Failed to require examinee to repeat sample		2	
Test 6: Visual Matching			
Failed to hold up book after sample on VM2	1		
Test 7: Numbers Reversed			
Fails to allow adequate time for responses		1	
Failed to prompt examinee to answer sample		1	
Gave example for incorrect start point		1	
Ceiling error		1	

All Errors by Examiner and Administration – Participant 8

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	10y2m	15y2m	15y6m
Verbatim error	9	11	6
Gave inappropriate feedback		1	
Did not record examinee response correctly	1	1	
Test 1: Verbal Comprehension			
Failed to request 1-word response			1
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	4	4	2
Failed to point to symbols when correcting	3	2	
Mispronounced stimulus word	1	2	1
Provided suffix on Story 7 unnecessarily	2		
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
Presented first item before giving prompt	1		
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback	1	1	1
Failed to query responses appropriately		1	
Asked student to repeat a non-sample item		1	
Incorrect ceiling		1	
Incorrect basal		1	
Test 6: Visual Matching			
Failed to hold up book after sample on VM2			1
Test 7: Numbers Reversed			
Fails to allow adequate time for responses		1	
Basal error	1		
Presented oral item prompts too fast or slow			2

All Errors by Examiner and Administration – Participant 9

	Administration		
	1	N/A	N/A
<i>Examinee's Reported Age</i>	6y7m		
Verbatim error	41		
Gave inappropriate feedback	6		
Did not record examinee response correctly	10		
Test 1: Verbal Comprehension			
Failed to request 1-word response	1		
Failed to query verbal response	1		
Failed to follow procedure if sample incorrect	1		
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	15		
Failed to point to symbols when correcting	1		
Did not require examinee to verbalize symbol	1		
Provided suffix on Story 7 unnecessarily	2		
Test 3: Visual-Spatial Processing			
Failed to provide feedback on sample	3		
Failed to ask "and what else?" if 2 pieces	1		
Test 4: Sound Blending			
Mispronounced stimulus word			
Test 5: Concept Formation			
Failed to acknowledge correct responses	2		
Failed to provide corrective feedback on errors	7		
Read instructions for incorrect start point	1		
Test 6: Visual Matching			
Failed to hold up book after sample on VM2	1		
Test 7: Numbers Reversed			
Gave example for incorrect start point	2		

All Errors by Examiner and Administration – Participant 10

	Administration		
	1	N/A	3
<i>Examinee's Reported Age</i>	8y9m		14y1m
Verbatim error	7		1
Gave inappropriate feedback	2		3
Did not record examinee response correctly	1		1
Test 1: Verbal Comprehension			
No errors			
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	18		
Failed to point to symbols when correcting	23		10
Did not query skipped symbol			1
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
No errors			
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback	1		
Test 6: Visual Matching			
Failed to hold up book after sample on VM2			1
Test 7: Numbers Reversed			
No errors			

All Errors by Examiner and Administration – Participant 11

	Administration		
	1	2	3
<i>Examinee's Reported Age</i>	16y9m	6y6m	14y11m
Verbatim error	4	25	6
Gave inappropriate feedback		6	3
Did not record examinee response correctly		10	1
Test 1: Verbal Comprehension			
Did not query verbal response	1		1
Mispronounced stimulus word			1
Failed to administer sample item			1
Queried clearly incorrect response		1	
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors		1	2
Failed to point to symbols when correcting	1	21	7
Mispronounced stimulus word	2	2	1
Provided suffix on Story 7 unnecessarily		3	
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
Did not remind to say smoothly		1	
Did not ask which of 2 responses was best		1	
Test 5: Concept Formation			
Failed to query responses appropriately		9	
Failed to acknowledge correct responses	4	3	
Test 6: Visual Matching			
Failed to hold up book after sample on VM2		1	1
Test 7: Numbers Reversed			
Fails to allow adequate time for responses	2		2
Gave prompt for incorrect start point	1		
Ceiling error			1

All Errors by Examiner and Administration – Participant 12

	Administration		
	1	N/A	N/A
<i>Examinee's Reported Age</i>	15y10m		
Verbatim error	4		
Gave inappropriate feedback			
Did not record examinee response correctly			
Test 1: Verbal Comprehension			
Ceiling error	1		
Test 2: Visual-Auditory Learning			
Mispronounced stimulus word	1		
Failed to correct an error	4		
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
Miscued audio track	1		
Mispronounced stimulus word	1		
Test 5: Concept Formation			
Failed to provide corrective feedback	1		
Failed to acknowledge correct responses	5		
Test 6: Visual Matching			
No errors			
Test 7: Numbers Reversed			
Failed to prompt examinee to answer sample	1		

All Errors by Examiner and Administration – Participant 14

	Administration		
	N/A	N/A	3
<i>Examinee's Reported Age</i>			16y4m
Verbatim error			5
Gave inappropriate feedback			
Did not record examinee response correctly			
Test 1: Verbal Comprehension			
No errors			
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors			2
Failed to point to symbols when correcting			6
Test 3: Visual-Spatial Processing			
Failed to give corrective feedback on sample			1
Test 4: Sound Blending			
Fails to allow adequate time for responses			1
Test 5: Concept Formation			
Provided feedback prematurely			1
Test 6: Visual Matching			
Failed to hold up book after sample on VM2			1
Test 7: Numbers Reversed			
Presented item backwards			1

All Errors by Examiner and Administration – Participant 15

	Administration		
	1	N/A	3
<i>Examinee's Reported Age</i>	11y8m		13y0m
Verbatim error	14		17
Gave inappropriate feedback	1		1
Did not record examinee response correctly	6		1
Test 1: Verbal Comprehension			
Did not query verbal response	1		
Basal error	2		
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors			1
Did not count extra word as error			1
Skipped line when correcting examinee	1		
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
Fails to allow adequate time for responses	1		
Failed to present additional sample items	1		
Test 5: Concept Formation			
Failed to query responses appropriately	1		
Test 6: Visual Matching			
Failed to hold up book after sample on VM2	1		
Test 7: Numbers Reversed			
Presented oral item prompts too fast or slow	7		

All Errors by Examiner and Administration – Participant 16

	Administration		
	1	N/A	3
<i>Examinee's Reported Age</i>	16y3m		11y4m
Verbatim error	16		8
Gave inappropriate feedback	3		3
Did not record examinee response correctly	2		
Test 1: Verbal Comprehension			
Queried clearly incorrect response			1
Test 2: Visual-Auditory Learning			
>6 sec before providing word on errors	1		
Corrected a correct response			1
Provided suffix on Story 7 unnecessarily	1		
Test 3: Visual-Spatial Processing			
No errors			
Test 4: Sound Blending			
No errors			
Test 5: Concept Formation			
Failed to use examinee's synonyms when providing feedback	1		1
Failed to query responses appropriately	1		
Failed to provide corrective feedback	1		1
Test 6: Visual Matching			
Failed to hold up book after sample on VM2	1		
Failed to prompt when stops at column end			1
Test 7: Numbers Reversed			
Presented item backwards	1		
Cued audio track incorrectly	1		
Failed to give sample item after reversal	1		
Basal error			1

Appendix E

Rationale for Classification of Common Errors as Systematically Affecting Examinee Scores

	Potential Effect on Scores	Rationale
Test 1: Verbal Comprehension		
Failed to query verbal responses	Decrease	Deprives examinees of an additional opportunity to correctly answer the item
Failed to request one-word responses	Decrease	Deprives examinees of an additional opportunity to correctly answer the item
Mispronounced stimulus word	Decrease	Examinees are less likely to recognize mispronunciations of stimulus words
Failed to establish ceiling	Decrease	Examinees do not have the opportunity to correctly answer items above the false ceiling
Failed to administer sample item	Decrease	Examinees do not have the benefit of sample items to learn test format
Queried an entirely incorrect response	Increase	Examinees have an additional opportunity to answer correctly when their incorrect response was not of adequate quality to warrant a query
Antonyms: Failed to ask for another response if “non”/”un”	Decrease	Deprives examinees of an additional opportunity to correctly answer the item
Failed to follow correct procedure if sample item incorrect	Decrease	Examinees do not have the benefit of correctly provided feedback
Failed to establish a basal	Decrease	Examinees' answers below the false basal are assumed correct
Test 2: Visual-Auditory Learning		
Failed to point to symbols when correcting	Decrease	Pointing helps establish a connection between the symbol and the provided word.
Mispronounced stimulus word (such as <i>the</i> , suffixes)	Decrease	Examinees may be confused by incorrectly presented stimuli

Failed to query skipped symbols	Decrease	Examinees do not have the opportunity to correctly answer skipped symbols; instead, they are marked as incorrect
Failed to correct an error	Decrease	Examinees do not have the additional opportunity to learn the correct association between symbol and word
Failed to count extra words as errors	Increase	Extra inserted words should be counted as errors, not ignored
Allowed examinee additional practice/time on new symbols	Increase	Examinees have more opportunity to form associations between symbols and words
Test 3: Spatial Relations		
Failed to give corrective feedback on sample items	Decrease	Examinees do not have the benefit of correctly provided feedback and may have incorrect answers inadvertently reinforced
Test 4: Sound Blending		
Failed to allow adequate time for responses	Decrease	Examinees may have provided better responses had the examiner paused the audio track and may feel expected to respond more quickly on subsequent items
Mispronounced stimulus word (in sample).	Decrease	Examinees may not understand the task if the sample item(s) do not form a proper blended word as pronounced by the examiner.
Test 5: Concept Formation		
Failed to use examinee's synonyms when providing feedback.	Decrease	Examinees are more likely to learn from feedback using their synonyms. The additional mental step of converting synonyms (e.g., from <i>circle</i> to <i>round</i>) should not be required
Failed to acknowledge correct responses	Decrease	Examinees are deprived of positive reinforcement for correct responses
Failed to provide corrective feedback on errors	Decrease	Examinees cannot learn from incorrect responses without corrective feedback
Improperly corrected an examinee's synonym	Decrease	Examinees may infer that their correct synonym is in error

Failed to ask examinee to repeat sample correctly	Decrease	Asking examinees to repeat the correct answer to sample items (after initially answering it incorrectly) reinforces the corrective feedback provided
Test 6: Visual Matching		
Failed to hold up book after sample	Increase	Examinees have additional time before the time limit to study test items
Test 7: Numbers Reversed		
Paused if examinee needs more time	Decrease	Examinees may have provided better responses had the examiner paused the audio track and may feel expected to respond more quickly on subsequent items
Failed to establish ceiling	Decrease	Examinees do not have the opportunity to correctly answer items above the false ceiling
Failed to give sample item after reversal for basal	Decrease	Examinees are deprived of exposure to an additional sample item
Failed to establish basal	Increase	Examinees' answers below the false basal are assumed correct

Luke Erichsen
luerichs@indiana.edu

Education

Ph.D. (2014), Indiana University, Bloomington, IN

Major: School Psychology (NASP approved) Minor: Counseling

Dissertation Title: *A Videotape Analysis of Examiner Error on the WJ-III COG*

M.S.Ed. (2012), Indiana University, Bloomington, IN

Major: Learning and Developmental Sciences – Educational Psychology Track

B.A. (2006), Whitman College, Walla Walla, WA, *magna cum laude*,

Phi Beta Kappa

Major: Religion

Minor: Psychology

Studied at the University of Otago, Dunedin, New Zealand, Spring 2005

Licensure

School Psychologist, Indiana Initial Practitioner License (expires 6/2016)

Predoctoral Internship

Integrated Behavioral Health Consortium of Indiana

Rotations in Muncie, Anderson, Elwood, and Alexandria, IN. August 2013 – July 2014

Supervisors: Sharon McNeany, Ph.D., HSPP and Linda Daniel, Ph.D., HSPP

Provide psychotherapy, assessment, and behavioral health consultation services in two primary care settings (IU Health Ball Memorial Hospital Family Medicine Residency Center and Madison County Community Health Center) as well as within two school corporations; coordinate care with physicians and allied health professionals; serve on multidisciplinary team providing follow-up care post hospital discharge and RTI prereferral team at an elementary school. Clients include children, adolescents, and adults.

Practicum Experience

The Project School

Bloomington, IN, Mar 2012 – Dec 2012

Supervisor: Thomas Huberty, Ph.D., ABPP

Conducted psychoeducational evaluations at a charter school with a special focus on serving a diverse population utilizing a project-based learning model. Most common areas of special education eligibility included specific learning disability and emotional disability.

<p>Madison County Community Health Center Anderson, IN and Alexandria, IN May 2012 – July 2013 <i>Supervisor:</i> Sharon McNeany, Ph.D., HSPP</p>	<p>Conducted individual therapy with children and adolescents in a Federally Qualified Health Center following an integrated care model. Primary duties involved conducting individual therapy and assessment in a rural intermediate school as part the center’s school-based clinic.</p>
<p>Damar Charter Academy Indianapolis, IN, Aug 2011 – May 2012 <i>Supervisor:</i> Jim Dalton, Psy.D., HSPP</p>	<p>Conducted psychoeducational and psychological evaluations at a charter school hosted by a residential treatment center and consulted with teachers on instructional planning. The school largely serves a residential population with developmental disabilities and extensive psychiatric histories. Most common areas of disability included cognitive (mild to severe), emotional/behavioral, and autism spectrum disorder. Gained extensive experience administering and interpreting instruments including the WISC-IV, SB5, WIAT-III, WJ-III ACH, ABAS-II, CBCL, and BASC-2.</p>
<p>Academic Well-Check Program Institute for Child Study Ellettsville, IN, Jan 2011 - May 2011 <i>Supervisor:</i> Rebecca Martínez, Ph.D., NCSP</p>	<p>Conducted and tracked curriculum-based measurements using AIMSweb system, designed individualized interventions targeting reading fluency and comprehension, and directly implemented interventions with students identified as academically at-risk.</p>
<p>Social and Behavior Support Program Institute for Child Study Bloomington, IN, Sep 2010 - Dec 2010 <i>Supervisor:</i> Russ Skiba, Ph.D.</p>	<p>Consulted with teachers in designing support plans for students with behavioral problems as part of a functional behavior assessment process. Provided one-on-one support in classroom to students.</p>
<p>South Central Community Action Program Catholic Charities Bloomington, IN, May 2010 - Aug 2010 <i>Supervisor:</i> Marsha McCarty, Ph.D., HSPP</p>	<p>Worked with preschoolers from low-SES backgrounds on improving social skills in a group setting.</p>

Forest Hills Special Education Cooperative Ellettsville, IN, Jan - May 2010
Supervisor: Amy Bartleson, Ed.S., NCSP

Assisted a school psychologist with professional activities, conducted psychoeducational evaluations, attended case conference committees.

Publications and Presentations

Erichsen, L. W., & Deskalo, A. (2014, February) *Disruptive mood dysregulation disorder: A new DSM-5 diagnosis for children*. Poster presented at the National Association of School Psychologists' 2014 Annual Convention, Washington, D.C.

Martínez, R. S., Floyd, R. & **Erichsen, L. W.** (2011). Strategies and attributes of highly productive scholars and contributors to the school psychology literature: Recommendations for increasing scholarly productivity. *Journal of School Psychology, 49*, 691-720.

Erichsen, L. W. (2011, February). *The academic school psychologist: A graduate student perspective*. Paper presented at the National Association of School Psychologists' 2011 Annual Convention, San Francisco, CA.

Erichsen, L. W., Jochim, M., White, S., Griddine, K., & Kirk, M. (2010, October). *Preservice training in Indiana's school psychology programs*. Paper presented at the Indiana Association of School Psychologists' 2010 Fall Conference, Indianapolis, IN.

Work Experience

Site Visitor

2/2012 – 5/2012 & 9/2012 – 5/2013
Center for Evaluation & Education Policy
Indiana University
Bloomington, IN

Conducted site visits in Indiana and Kentucky for an evaluation of the 21st Century Community Learning Centers, a federal grant program for afterschool programs in schools serving students at-risk.

Graduate Assistant

8/2012 – 7/2013
Office of the Dean
Indiana University School of Education
Bloomington, IN

Provided administrative support to the Executive Associate Dean. Responsibilities included attending and writing minutes for the Policy Council, preparing advertisements for faculty openings, updating the website for

Associate Instructor

8/2010 – 5/2012
Indiana University School of Education
Bloomington, IN

Taught three sections of G203, a communication and counseling skills course for education majors. Also taught two online sections of P248, a child development course for

<p>Adjunct Instructor 9/ 2008 – 5/2009 Montana State University – Billings Billings, MT</p>	<p>Taught three sections of English 100 (English Essentials), a course designed for developmental- level writing students. Tutored all levels of writing students at the university’s Academic Support Center.</p>
<p>Counselor 7/ 2008 – 8/ 2009 Yellowstone Boys and Girls Ranch Billings, MT</p>	<p>Provided direct care support to adolescents with emotional and behavioral disorders in a long-term, secure residential treatment setting. Led informal groups and provided individual support. Trained in Therapeutic Crisis</p>
<p>Institutional Research Intern 9/2005 – 5/ 2006 Whitman College, Walla Walla, WA</p>	<p>Prepared reports for college administrators and the National Survey of Student Engagement, analyzed data, updated the Office of Institutional Research’s website, and interpreted survey results.</p>

Service

- Co-President, Student Affiliates in School Psychology, Indiana University, 2010-2011
- Recruitment Chair, Student Affiliates in School Psychology, Indiana University, 2009-2010
- Full-time volunteer for religious organization, 2006-2008

Awards and Honors

- Indiana University School of Education Fellowship, 2009-2013
 - Most selective fellowship awarded to incoming doctoral students
- Phi Beta Kappa, Whitman College, 2006
- Honors in Major, passed oral defense of thesis “with distinction,” Whitman College, 2006

Professional Memberships

- National Association of School Psychologists (NASP)
- Indiana Association of School Psychologists (IASP)
- American Psychological Association (APA) Division 16