A COMPARISON OF THE WISC-V PROCESSING SPEED SUBTESTS: PAPER-PENCIL AND DIGITAL FORMATS

A DISSERTATION

SUBMITTED TO THE FACULTY

OF

THE GRADUATE SCHOOL OF APPLIED AND PROFESSIONAL PSYCHOLOGY

OF

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY

BY

AMANDA MARIE FERRIOLA

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

OF

DOCTOR OF PSYCHOLOGY

NEW BRUNSWICK, NEW JERSEY AUGUST 2019

APPROVED: ________________________
Ryan Kettler, PhD

____
Kenneth Schneider, PhD

DEAN: ________________________
Francine Conway, PhD

ABSTRACT

In recent years, there has been a movement toward use of digital platforms to administer cognitive, achievement, and neuropsychological assessments. Principal changes were made to the Processing Speed subtests of the Wechsler Intelligence Scale for Children, Fifth edition (WISC-V) - Coding (CD) and Symbol Search (SS) - for adaptation from a paper-pencil format to a digital format. The aim of the current study is to assess whether paper-pencil and digital formats of CD and SS interchangeably measure the same construct: processing speed. The impact of psychomotor coordination on differences in performance between paper and digital versions of subtests was also examined. A total of 41 students between the ages of 13.0 and 16.11 were administered a demographic questionnaire and the Beery-Buktenica Developmental Test of Visual-Motor Integration – Sixth Edition (Beery VMI) full form followed by CD in digital format (CD-D), CD in paper-pencil format (CD-P), SS in digital format (SS-D), and SS in paper-pencil format (SS-P) administered in counterbalanced sequences. Data was analyzed using descriptive statistics, Pearson correlations, and a two-tailed paired t-test to assess relationships between variables. Results of the study indicate correlations between paper-pencil and digital scores on CD, SS, and the Processing Speed Index (PSI) were lower than hypothesized thresholds for equivalence. The Multitrait Multimethod Matrix (MTMM) was applied to examine evidence for convergent and discriminant validity of CD and SS; results were mixed. Finally, examinees' scores on measures of visual-motor coordination did not share a significant relationship with differences between scores on the digital format and on the paper-pencil format. Limitations of the study, as well as implications for practice and future research directions, are discussed.

ACKNOWLEDGEMENTS

As I reflect on the process of completing a dissertation, I am reminded of similar feats of endurance climbing mountains and running marathons. Every great adventure begins and ends with a single step. Every great adventure includes an invaluable community tasked with reminding the adventurer to continuously place one foot in front of the other and to mind her path. With the conclusion of my dissertation and in turn my doctoral degree I am afforded the opportunity to reflect on my community, without whom I would have undoubtedly lost my way. First, to my advisor and dissertation chair, Ryan Kettler, from whom I learned about formal writing, research methods, professionalism, and the field of psychology. I would like to express gratitude to my committee member, Ken Schneider, for inspiring me to approach each situation with kindness, compassion, and a healthy dose of skepticism. To Michael Friedman, partly for affording me access to needed materials for data collection, and mostly for being a wonderful human and an even better friend. To Docia Demmin, for her endless patience and willingness to provide her support and wisdom. To Docia, Jaye Odom, Jacqueline Shinall, and Laura Lesnewich for lending me their time and assessment skills so selflessly. Finally, and most of all, I wish to thank my husband and family for fostering an environment of immeasurable love, support, and humor. I am forever indebted to you; this is as much your accomplishment as mine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

Literature Review

Introduction

In recent years, there has been a movement toward use of digital platforms to administer cognitive, achievement, and neuropsychological assessments (Daniel, 2012a; Daniel, 2012b; Daniel, 2012c; Daniel, 2013; Daniel, Wahlstrom, & Zhang, 2014; Daniel, Wahlstrom, & Zhou, 2014; Raiford, Holdnack, Drozdick, & Zhang, 2014; Raiford, Drozdick, & Zhang, 2015; Raiford, Drozdick, & Zhang, 2016; Raiford & Zhang et al., 2016). The Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V), a standardized measure of cognitive functioning, was released in 2014 (Wechsler, 2014). At the time of release, all subtests were available in paper-pencil and digital formats, except for the Processing Speed Index (PSI) subtests, which were only available with use of paper-pencil response booklets. In 2016, two years after the release of the WISC-V, Pearson adapted the PSI subtests for administration in digital format (Raiford & Zhang et al., 2016). This was the first attempt at adaptation of a Wechsler processing speed assessment to digital format without the aid of physical manipulatives.

The aim of the current study is to assess whether paper-pencil and digital formats of processing speed subtests from a cognitive assessment (WISC-V) interchangeably measure the same construct: processing speed. Wechsler processing speed scores reflect an examinee's "speed and accuracy of visual identification, decision making, and decision implementation" and performance on such measures is "related to visual scanning, visual discrimination, short-term visual memory, visuomotor coordination, and concentration" (Wechsler, 2014).

Processing Speed

Processing speed, generally, is not easily definable; researchers in the field are not able to confidently isolate the fundamental cognitive mechanisms that underlie the construct (Shanahan et al., 2006). This makes uniform measurement of processing speed challenging, as it is unclear whether scores from the measurement tool reflect the intended construct. Processing speed has been measured as a component of the larger concept of fluid intelligence (Cattell, 1963; Keith & Reynolds, 2010). Fluid intelligence, an aspect of intelligence and a critical factor in learning, is the ability to solve novel problems using reason instead of preexisting knowledge (Sattler, 2008).

Processing speed definitions. Although definitions of processing speed are not uniform, they do have common characteristics. In a study assessing the role of processing speed in Attention Deficit and Hyperactivity Disorder (ADHD) and Reading Disabilities, results showed processing speed measures tapped into two separate and correlated factors: rapid naming of stimuli and speeded motor and non-linguistic abilities (Shanahan et al., 2006). Jacobson et al. (2011) postulate components of processing speed include sensory registration and perception, response preparation, and response execution. Response preparation generally includes aspects of visual reprocessing, mental transformation, associative memory, and response selection. Response execution, specifically in the context of measurement of processing speed, generally includes a graphomotor response component (Jacobson et al., 2011). Horn and Blankson (2005) define processing speed as an ability to rapidly scan and react to simple tasks. A unifying summary of processing speed definitions is: speeded sensory input and subsequent output during simple tasks. Wechsler (2014) emphasizes decision making is an important aspect of processing speed because it distinguishes processing speed from simple reaction time.

Processing speed measurement. Psychological measurement may be used to meaningfully convey information about concepts such as processing speed (Sattler, 2008). To measure a concept, certain requirements must be met. Sattler (2008) outlines four requirements of measurement. To begin, a concept must be identified and clearly defined. A method of measurement must be chosen (e.g., a tool or operation) that is compatible with the concept being measured. The rules of measurement, specific to the identified concept and measurement method, must be outlined. Finally, results must be expressed as units on a scale to communicate meaning (Sattler, 2008). A norm-referenced measurement provides a relevant normative group for comparison of results (Sattler, 2008). Comparisons may be made after raw scores are converted into a derived score. Derived scores allow for comparison to other peers and comparison to other measures completed by the same examinee. Norm-referenced measurement requires administration and scoring procedures to be standardized (Sattler, 2008). Absent use of norm-referenced measurement, an examinee's raw score is meaningless, because no context is provided for examinee performance relative to various characteristics including age and level of education. Norm-referenced assessments, such as the WISC-V, are used to measure processing speed. Results can be compared to same aged peers to provide contextually meaningful information about an examinee's processing speed performance.

A method used to compare one norm-referenced assessment to another is to calculate a Pearson correlation. Cohen (1988) developed guidelines for determining the strength of a correlation coefficient based on its magnitude (See Table 1.) These guidelines will be used throughout the document to discuss the strength of Pearson correlation coefficients.

Table 1.

Guidelines to determine strength of Pearson correlation coefficient

Coefficient Value	Strength of Association
0.1 < | *r* | < .30	Small correlation
.30 < | *r* | < .50	Moderate correlation
| *r* | < .50	Strong correlation

Adapted from Cohen (1988)
| r | indicates the absolute value

Attempts to include processing speed as a measure of intelligence began as early as 1884 with Sir Frances Galton (Sattler, 2008). He opened a psychometric laboratory to the public at the International Health Exhibition. Galton used objective techniques to measure intelligence through sensory discrimination abilities. He hypothesized geniuses among the population would have the most superior sensory abilities, because human knowledge is acquired through the senses. Although this hypothesis was later debunked, Galton was the first researcher to include speed of information processing as a measure of intelligence (Sattler, 2008). James Mckeen Cattell worked as an assistant to Galton prior to opening a psychology laboratory at the University of Pennsylvania. He coined the term “mental test” in 1890 and compiled a battery of tests for evaluating skills. Among them were measures of Rate of Movement, Reaction-Time for Sound, and Time for Naming Colors. While these skills were eventually judged to be poor measures of cognitive ability, they were considered early measurements of speed of information processing and opened the door for the empirical study of intelligence (Sattler, 2008). Reaction time and inspection time are two accepted measurements of speed of information processing. Reaction time as a measureable component of speed of information processing can be further dichotomized to simple reaction time and complex reaction time. Two types of complex reaction time are recognition and choice reaction time. See Table 2 for definitions of aforementioned constructs.

Table 2.

Measurement of Speed of Information Processing

i. *Reaction Time (RT)* - time between exposure to a stimulus and motor response to stimulus
    a. Simple RT – time between onset of stimulus and reaction
    b. Complex RT – time between presentation of one or more stimulus and differential response
        i. Recognition (Go/No-Go) – Examinee is required to respond to presentation of one stimulus and inhibit response to another stimulus
        ii. Choice RT (CRT) – Examinee is required to respond in a certain way to presentation of one stimuli and respond in a different way to another stimulus

ii. Inspection Time – speed of encoding, or the minimum amount of time needed to discriminate between two or more stimuli

Adapted from Sattler (2008).

Speed of information processing measured by inspection time correlates with IQ at about *r* = -.29 for children (small correlation). Higher IQ is associated with faster information processing (Sattler, 2008). Processing speed has limitations as a measure of intelligence, including variability in test-retest reaction time. Reaction times collected from individuals taking the same processing speed measure five days apart will correlate at *r* = .60 (strong correlation; Sattler, 2008).

Digital Test Administration

A recent trend in psychological assessments, including measures of processing speed, has been movement away from paper-pencil administration and toward digital administration (Daniel, 2012a; Daniel, Wahlstrom & Zhang, 2014; Raiford & Zhang et al., 2016). Q-interactive is a digital platform developed by Pearson to enable examiners to score and administer individual tests with computer assistance (Daniel, 2012a; Daniel, Wahlstrom & Zhang, 2014; Raiford & Zhang et al., 2016). To demonstrate equivalence between paper and digital versions of individual tests, Pearson developers released a series of Q-interactive Technical Reports (Daniel, 2012a;

Daniel, 2012b; Daniel, 2012c; Daniel, 2013; Daniel, Wahlstrom, & Zhang, 2014; Daniel, Wahlstrom, & Zhou, 2014; Raiford et al., 2014; Raiford et al., 2015; Raiford & Drozdick et al., 2016; Raiford & Zhang et al., 2016). Studies were compiled on cognitive, academic, and language measures. Equivalence studies of cognitive measures included assessments such as the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Daniel, 2012a), the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Daniel, 2012b), the California Verbal Learning Test - Second Edition (CVLT-II, Daniel, 2012c), select subtests from the Delis-Kaplan Executive Function Scale (D-KEFS; Daniel, 2012c), and the WISC-V (Daniel, Wahlstrom, & Zhang, 2014; Raiford & Zhang et al., 2016). Academic tasks included selected subtests from measures such as the Wechsler Individual Achievement Test – Third Edition (WIAT-III; Daniel, 2013). Language tasks included measures such as the Clinical Evaluation of Language Fundamentals – Fifth Edition (CELF-5; Daniel, Wahlstrom, & Zhou, 2014). Special group studies were conducted on populations including children with Autism Spectrum Disorders, Learning Disabilities, Language or Hearing Impairments, Attention Deficit/Hyperactivity Disorders, Intellectual Giftedness, and Intellectual Disabilities (Raiford et al., 2014; Raiford et al., 2015; Raiford & Drozdick et al., 2016; Raiford & Zhang et al., 2016).

Effects of digital administration. Daniel and colleagues (Daniel, 2012a; Daniel, Wahlstrom, & Zhang, 2014) highlighted various ways digital administration of tests in general may affect scores. The first involves the examinee interaction with the tablet. The second involves the examiner interaction with the tablet. The third involves the digital administration environment.

Examinee-tablet interaction. Throughout digital administration, two tablets, one for the examinee and one for the examiner, are synched. Not all tasks require the use of the examinee tablet for responses. For example, participants respond orally on most verbal measures. Whenever both tablets are in use, the examinee can view visual stimuli and respond to items by touching (or tapping) the screen of the tablet. Prior use of, or exposure to, a tablet is not required. It is possible familiarity with a tablet may increase an examinee's comfort with this format.

Examiner-tablet interaction. The examiner can use a stylus or her finger to manipulate the tablet's screen. The examiner's tablet displays the examinee's screen in real-time, allowing the examiner to manipulate examinee responses and control the content displayed on the examinee's screen. The examiner can view administration instructions and prompts throughout administration. Additionally, the examiner's tablet records time and examinee touch and audio responses. At any point during administration, the examiner can access a digital note pad to record notes linked to specific items, specific subtests, or the overall assessment.

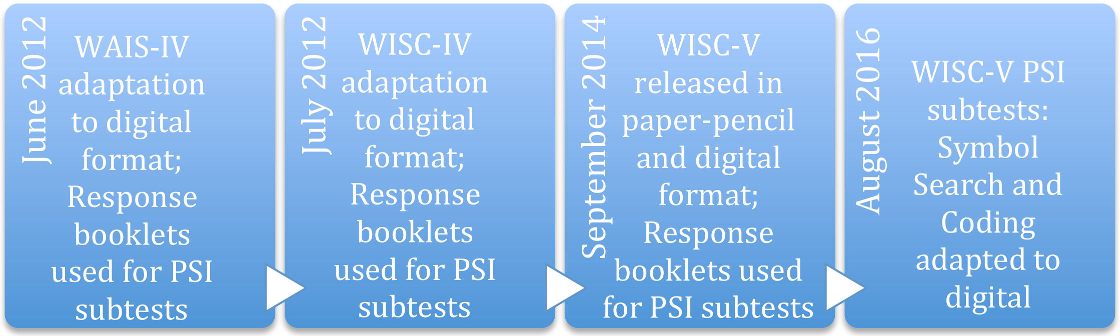
Digital administration environment. The digital administration environment includes the global effects of digital tools during administration of an assessment. In the context of equivalence studies, the general goal of digital administration is to equate results with results from correct paper administration. Non-equivalent digital results due to error were discovered through video recordings of administration and, when possible, corrected. For instance, video recordings revealed a pattern in which participants tended to give shorter verbal responses to examiners whom appeared to have difficulty rapidly typing responses on an early digital version of a Wechsler subtest involving an examiner keyboard (Daniel, 2012b). The Q-interactive team determined this affected the accuracy with which participants were providing responses and so eliminated the use of keyboards during administration. Another example is the accuracy with which the digital system captures examinee responses provided by touching the tablet screen compared to manual recording of responses. Many early versions of digital administration required the examiner to enter a score rather than have the tablet automatically record scores for touch responses. Q-interactive is moving toward the latter feature to aid the examiner in administration (Daniel, 2012a).

Physical Manipulatives in Digital Administration

Physical manipulatives include items provided to an examinee with which she is asked to interact. There is a psychomotor component inherent to items administered with physical manipulatives. This is particularly true for subtests administered using response booklets. For paper-pencil administration of Coding, participants do not earn points for items replicated with poor formation (Wechsler, 2014). Adequate psychomotor and coordination (e.g., the ability to grip a pencil) capabilities, therefore, are assumed prerequisites to accurate measurement of processing speed based on paper-pencil subtests. Prerequisite, or access skills, are skills a participant is required to use to take the test. For example, a third grade student taking a science test may be required to read a paragraph in order to provide an answer to a question. In this instance, the construct being measured is his or her knowledge of science; the access skill is his or her ability to read. His or her responses may not accurately reflect his or her knowledge of science if he or she has a difficult time reading. With regard to processing speed, fine motor precision is relevant for tests requiring it as an access skill. An argument could be made that a more accurate measurement of processing speed involves exclusion of access skills such as fine motor precision. This becomes possible with digital administration. The ability to accurately reconstruct a symbol with fine motor precision should only be an access skill in measurement of the cognitive concept of processing speed if it is consistently measured across all formats.

Within the context of test administration, physical manipulatives include such items as physical response booklets, blocks, and grids (Daniel, 2012a; Daniel, Wahlstrom, & Zhang, 2014). In early versions of digitally administered assessments, as previously discussed, physical manipulatives were used. Digital adaptation of physical manipulatives was considered a potential threat to equivalence. Ultimately, the hope was to determine the function of physical manipulatives in the context of the Wechsler intelligence assessments for adaptation to a digital format. Daniel (2012a) stated, “though these physical components may eventually be replaced by interactive digital interfaces, the degree of adaptation required would make raw-score equivalence unlikely” (p.1). Therefore, during adaptation, the approach switched from the demonstration of equivalence to the demonstration of reliability and validity of newly adapted digital scores.

In June 2012, the WAIS-IV subtests were adapted to the Q-interactive (digital) platform by demonstrating raw score equivalence between paper-pencil and digital administrations and applying preexisting paper-pencil norms. Regardless of test format, physical manipulatives continued to be used for processing speed (response booklets) and working memory (blocks) subtests. In July 2012, the WISC-IV underwent a similar adaptation to the Q-interactive (digital) platform. An equivalence study was conducted and the norms collected for the paper-pencil format were applied to the digital format. Again, response booklets and blocks were used for processing speed and working memory subtests, respectively. In September 2014, the WISC-V was released. In line with its predecessors, equivalence was demonstrated between paper-pencil and digital formats and response booklets and blocks were used for processing speed and working memory subtests, respectively. In August 2016, WISC-V processing speed subtests, Coding and Symbol Search, were adapted to the Q-interactive platform. Changes to the subtests during adaptation were substantial, therefore, new data were collected and a scaling procedure was applied to administration with response booklets and administration with iPads only. The examiner currently has the ability to choose to administer Coding and Symbol Search with or without response booklets, as the formats are used interchangeably (See Figure 1).


June 2012
WAIS-IV adaptation to digital format; Response booklets used for PSI subtests
July 2012
WISC-IV adaptation to digital format; Response booklets used for PSI subtests
September 2014
WISC-V released in paper-pencil and digital format; Response booklets used for PSI subtests
August 2016
WISC-V PSI subtests: Symbol Search and Coding adapted to digital

Note. Compiled from Daniel 2012a; Daniel 2012b; Daniel, Wahlstrom, and Zhang, 2014; Raiford and Zhang et al., 2016
Figure 1. Timeline: Wechsler Processing Speed (PSI) Subtests

It is interesting to note, to date, a number of digital formats continue to use physical manipulatives. For example, the WAIS-IV processing speed subtests (i.e., Symbol Search, Coding, and Cancellation), the WISC-V processing speed subtest, Cancellation, and the D-KEFS Trail Making Test continue to be administered using printed response booklets (Daniel, 2012a; Daniel, 2012c; Daniel, Wahlstrom, & Zhang, 2014; Raiford & Zhang et al., 2016). Additionally, Block Design subtests from the WAIS-IV and the WISC-V continue to require administration with physical blocks (Daniel, 2012a; Daniel, Wahlstrom, & Zhang, 2014).

Considerations and Correlations between Measures of Processing Speed

Researchers involved in Q-interactive equivalence studies consistently chose an effect size of less than 0.2 (Cohen's d) as the standard for equivalence. This standard is questionable because of the range of scores that are considered equivalent. Wechsler measures have a mean of

10 and a standard deviation of 3. Theoretically, a subtest score yielding means 0.6 of a scaled score point higher or lower than another subtest score can be considered equivalent to it.

It is important to consider the correlations between index scores and subtests scores on various assessments. Throughout equivalence studies, Pearson uses an *r* of .70 as the threshold for application of equating procedures. Index scores measuring the same construct are expected to correlate highly. The correlation between two assessments intended to measure the same construct is assumed to be higher than the correlation between two assessments intended to measure different constructs. The reliability of processing speed subtests can be measured using a test-retest reliability coefficient (Wechsler, 2014). Test-retest reliability coefficients for Wechsler processing speed measures are between .83 and .87. A measure will not correlate higher with another measure than it does with itself within a short timeframe. This is true across Wechsler measures because the PSI correlations range from *r* = .34 (moderate correlation) for WISC-V paper-pencil and WPPSI-IV and *r* = .79 (strong correlation) for WISC-V paper-pencil and WAIS-IV (see Table 3).

Processing speed scores on the WAIS-IV, WISC-IV, WISC-V, and WPPSI-IV correlate strongly with each other. Processing speed scores on Wechsler measures (WAIS-IV, WISC-IV, WISC-V, and WPPSI-IV) correlate higher with each other than with index scores on the Woodcock Johnson Test of Cognitive Abilities – Fourth Edition (WJ-COG IV) because Wechsler task were more similar to each other than tasks from WJ-COG IV. The Multitrait Multimethod Matrix (MTMM) is used to examine evidence for convergent and discriminant validity of measures (Campbell & Fiske, 1959). Within MTMM, variance in each measure is partly due to the construct and partly due to the method. It is assumed scores on assessments that share a construct and use different methods correlate higher than scores on assessments that share a method and measure different constructs. Among scores on assessments that measure the same construct, scores intended for use interchangeably (i.e., scores from the same assessment administered in a different format) correlate the highest. The Processing Speed Index (PSI) scores on the WISC-V-P are expected to correlate the highest with the PSI scores on the WISC-V-D, though this information is not included in equivalence studies; the current study addresses this relationship.

Table 3.

Processing Speed Index Correlations

	WISC-V (paper)	WPPSI-IV	WISC-IV	WAIS-IV	WJ- COG IV
WISC-V (paper)	**(.83)**	.34	.70	.79	-
WPPSI-IV	.34	**(.84)**	.56	-	-
WISC-IV	.70	.56	**(.86)**	.68	.55
WAIS-IV	.79	-	.68	**(.87)**	.44
WJ-COG IV	-	-	.55	.44	-

Note. Compiled from Wechsler 2014 and McGrew, LaForte and Schrank, 2014.
Correlations between tests

Test-retest reliability coefficients

Subtest scores are expected to follow the same pattern as index scores. Scores on Wechsler subtests measuring the same construct are expected to correlate. Scores on measures intended for use interchangeably are expected to correlate the highest. For example, Coding (CD) from the WISC-V Paper-pencil (WISC-V-P) is expected to correlate higher with CD on the WISC-V Digital (WISC-V-D) than with CD on the WAIS-IV because CD from the WISC-V-P is used interchangeably with CD on WISC-V-D. In actuality, the correlations between CD on WISC-V-P and CD on WAIS-IV (*r* = .69) and CD on WISC-V-P and CD on WISC-IV (*r* = .69) are higher than the correlation between CD on WISC-V-P and CD on WISC-V-D (*r* = .63). Symbol Search (SS) correlations are consistent with the expected pattern. Correlations between SS on WISC-V-P and SS on WAIS-IV (*r* = .61) and SS on WISC-V-P and SS on WISC-IV (*r* =

.54) are lower than the correlation between SS on WISC-V-P and SS on WISC-V-D (*r* = .67; see Table 4).

Table 4.

Coding and Symbol Search Correlations

	Assessments	Coding (*r*)	Symbol Search (*r*)
Correlations between tests	WISC-V (paper)/WISC-V (digital)	.63	.67
	WISC-V (paper)/WAIS-IV (paper)	.69	.61
	WISC-V (paper)/WISC-IV (paper)	.69	.54
Test-retest reliability coefficients	WISC-V (paper)	.81	.80
	WISC-V (digital)	.80	.75
	WISC-IV (paper)	.87	.78
	WAIS-IV (paper)	.86	.81

Note. Compiled from Wechsler 2014.

Establishing Equivalence

In the current example, processing speed is the construct being measured. Performance on digital assessment and performance on paper-pencil assessment are the units that can be equated. Assessments were developed with the intent they would yield equivalent scores by converting raw scores to the same metric. Each measure is slightly more or less precise than the next; therefore, perfectly interchangeable units are not attainable in practice. There are several limitations to establishing equivalence between the two measures. Dorans (2004) indicated to use scores interchangeably; certain prerequisites are obligatory prior to applying equating methods. First, adequate reliability is necessary for each format (i.e., digital and paper-pencil). Score equating should only occur if it is determined that both versions reliably measure the same construct (Dorans & Holland, 2000). Digital and paper-pencil assessments should each measure the construct they claim to measure (i.e., processing speed). It is necessary the correlation between raw scores from both formats is high to apply equating procedures. Finally, the aforementioned prerequisites must be true and consistent across different subpopulations (Dorans, 2004). Relevant variables in measurement of processing speed include sex, ethnicity, geographic region, and level of parental education. In addition, in a scenario in which assessments A and B are both administered, it is possible that assessment A may yield higher scores than assessment B simply because assessment A is easier. In this instance, use of a conversion technique is required, such as collecting separate norms for each assessment (Kolen & Brennan, 2014).

Paper-Pencil and Digital Equivalence Designs

Q-interactive equivalence studies used four designs to evaluate equivalence: (a) a randomly equivalent groups design, (b) a non-randomly equivalent groups design, (c) a test-retest design, and (d) a dual capture design (Daniel, 2012a; Daniel, 2012b). In equivalent-groups designs, a sample size larger than that needed for retest and dual capture designs is required because each examinee takes each subtest one time (as opposed to taking each test more than one time as in a retest design). The form of each subtest (i.e., paper-pencil or digital) will vary depending on the condition to which the examinee is assigned. This design is preferable for subtests that have practice effects; it eliminates unwanted effects of taking a subtest more than one time. Equivalent-groups designs were used most frequently to avoid practice effects (Daniel, 2012a). After an examinee has been exposed to a problem, practice effects refer to strategies to more effectively approach said problem. In equivalent-groups designs, practice effects are avoided because a group of participants taking the test in digital format is compared to a group of participants taking the test in paper-pencil format; examinees only take each test one time. Equivalent-groups designs are assigned either randomly or non-randomly.

(a) Randomly equivalent groups design. In random assignment, the sample resembles the general population with regard to sex, ethnicity, and level of education; age components are dependent on the research question (e.g., older participants are overrepresented because digital administration is hypothesized to have a larger effect on that population). Participants are randomly assigned to a format: digital or paper-pencil. Participants are administered covariate tests in paper-pencil format immediately following focal test (i.e., the test(s) of primary relevance to the study) administration. Covariate tests measure the same construct(s) as the focal test. For example, Daniel (2012b) looked at format effects (digital vs. paper-pencil) of WISC-IV standard battery. In this example, the focal test is the WISC-IV. Covariate tests included measures of verbal and nonverbal intelligence, processing speed, and working memory taken from assessment batteries similar to the WISC-IV (See Table 5).

Table 5.

Randomly Equivalent Groups Design: Subtest Administration Example

Tests/Subtests	Participant A (Digital Format)	Participant B (Paper-Pencil Format)
Block Design	**Digital**	**Paper**
Similarities	**Digital**	**Paper**
Digit Span	**Digital**	**Paper**
Matrix Reasoning	**Digital**	**Paper**
Vocabulary	**Digital**	**Paper**
Arithmetic	**Digital**	**Paper**
Symbol Search	**Digital**	**Paper**
Visual Puzzles	**Digital**	**Paper**
Information	**Digital**	**Paper**
Coding	**Digital**	**Paper**
Letter-Number Sequencing	**Digital**	**Paper**
Figure Weights	**Digital**	**Paper**
Comprehension	**Digital**	**Paper**
Cancellation	**Digital**	**Paper**
Picture Completion	**Digital**	**Paper**
Kaufman Brief Intelligence Test, Second Edition*	*Paper*	*Paper*
Speed of Information Processing**	*Paper*	*Paper*
Letter Span***	*Paper*	*Paper*

***KBIT-2: Yields Verbal and Nonverbal ability scores**
****Subtest from the Differential Ability Scales, Second Edition (DAS-II)**
*****Subtest from the WISC-IV Integrated**
Focal test are presented in bold.
Covariate test are presented in italics.
Note. Developed from Daniel (2012b).

Multiple regression or ANCOVA is used to separately analyze results of focal and covariate tests. The focal test (with an age-adjusted normative score) serves as the dependent variable. Predictor variables are demographic variables (sex, ethnicity, level of education), covariate tests (with age-adjusted normative scores), and format (digital or paper-pencil). Format effect is the unstandardized regression weight for the dummy variable (Daniel, 2012a). Effect size is calculated by dividing the average format effect by 3 (i.e., the standard deviation of the focal test's normative-score metric).

(b) Non-randomly equivalent groups design. The non-randomly equivalent groups design is similar to the randomly assigned equivalent groups design except for the use of preexisting norms. This design makes use of preexisting norms from paper-pencil administrations; norms from a demographically similar population are collected with digital administration as the focal test and paper-pencil covariates. Covariates in the non-randomly equivalent groups design are subtests administered to the norm sample and the newly collected sample in paper-pencil format. For example, Daniel (2012a) assigned participants to one of two groups. Each group was administered the entire WAIS-IV battery: half in digital format (focal subtests) and half in paper-pencil format (covariates). Subtests were assigned to a format based on domain; the goal was to divide each domain as evenly as possible to represent both format and construct. To illustrate the point: Coding, Symbol Search, and Cancellation are subtests that measure processing speed. If Participant A is assigned to Group 1, she is administered Coding and Symbol Search in digital format and administered the covariate, Cancellation, in paper-pencil format. If Participant B is randomly assigned to Group 2, he is administered Cancellation in digital format and Coding and Symbol Search, the covariates, in paper-pencil format (See Table 6). Prediction equations are developed based on demographics and covariate-test scores

(Daniel, 2012a). Prediction equations allow researchers to use existing data to estimate scores of demographically similar populations (McGrew, LaForte & Schrank, 2014). Format effects are the difference between the focal tests' observed (i.e., actual scores) and predicted scores. This design requires a smaller number of participants than the randomly equivalent groups design because it uses data that has already been collected from the norm sample. It does not have the benefits of random assignment in that participants are not equivalent across all characteristics, both identified and unknown, having the potential to influence test performance (Daniel, 2012a).

Table 6.

Non-Randomly Equivalent Groups Design: Subtests Administration Example

Subtest Order	Format (Group 1)	Format (Group 2)
Block Design	*Paper*	**Digital**
Similarities	*Paper*	**Digital**
Digit Span	**Digital**	*Paper*
Matrix Reasoning	**Digital**	*Paper*
Vocabulary	**Digital**	*Paper*
Arithmetic	*Paper*	**Digital**
Symbol Search	*Paper*	**Digital**
Visual Puzzles	**Digital**	*Paper*
Information	**Digital**	*Paper*
Coding	*Paper*	**Digital**
Letter-Number Sequencing	*Paper*	**Digital**
Figure Weights	*Paper*	**Digital**
Comprehension	*Paper*	**Digital**
Cancellation	**Digital**	*Paper*
Picture Completion	**Digital**	*Paper*

Focal tests are in bold.
Covariate tests are in italics.
Note. Developed from Daniel 2012a.

(c) Retest design. In retest designs, the digital and paper-pencil formats of each test are administered to the same examinees. In other words, each examinee takes the same test twice in each format, and serves as his or her own control. Subtests are administered in a counterbalanced sequence, meaning half of the participants take the digital format first and half of the participants take paper-pencil format first. For example, in Group 1, participants take Coding and Symbol Search in digital format followed by Coding and Symbol Search in paper-pencil format. Group 2 participants are administered the same subtests in paper-pencil format followed by digital format (See Figure 2). To determine impact of format effect, the difference between the second administration score and first administration score is calculated. If a format effect exists, half of the difference scores will increase and half will decrease, depending on sequence. If no format effect exists, sequence will not affect performance and the average value of differences will be equal. This design is used whenever practice effects are not anticipated and response processes are unlikely to change (Daniel, 2012a). It is preferable to use, whenever possible, because it requires a smaller number of participants compared to equivalent-groups designs. Drawbacks include examinee time commitment because each examinee is required to take each test two times. Studies evaluating processing speed use this design most frequently because subtests contain non-meaningful stimuli and improvement from strategy use (i.e., practice effects) is minimal (Daniel, 2012a).

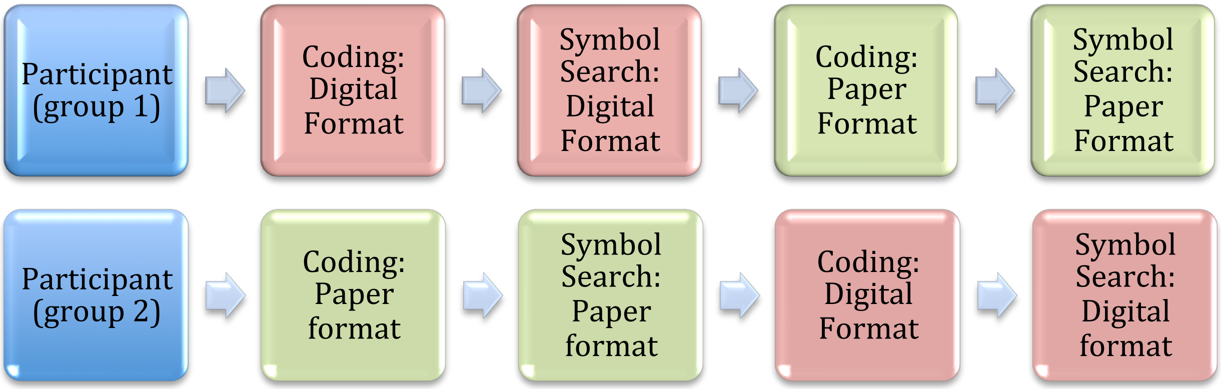


Figure 2. Retest design: subtest administration example. Developed from Daniel, 2012a

(d) Dual capture design. The dual capture design is used exclusively to assess the effect of digital administration on the examiner's ability to score and capture examinee performance. This design is used whenever format is expected to affect the examiner's experience and the examinee's experience is expected to remain the same as during paper-pencil administration. All measures are administered to each examinee one time and each examinee's responses are video recorded from the examiner's perspective. Examiners are randomly assigned to digital or paper-pencil format, then watch the video administration and record participants' responses in the assigned format (Daniel, 2012a; Daniel, Wahlstrom, & Zhang, 2014). Scores in digital format are averaged and scores in paper-pencil format are averaged. The averages are subsequently compared to assess the impact of format effects on the scoring process. The examiner engages in little, if any, recording on processing speed subtests as participants record their own responses in both formats. Therefore, this design is not particularly relevant for processing speed subtests.

Paper-Pencil and Digital Equivalence Studies

For Wechsler equivalence studies, the standard effect size for equivalence was set at 0.2. An effect size describes the size of the difference between two variables. "An effect size of 0.2 is slightly more than one-half of a scaled-score point on the commonly used subtest metric that has a mean of 10 and standard deviation of 3" (Daniel, 2012a).

WAIS-IV. The WAIS-IV was first adapted to the Q-interactive platform in 2011 (Daniel, 2012a). It was the first Wechsler scale to be adapted into digital format. The goal was to demonstrate equivalence between existing raw scores on the paper-pencil format and newly collected scores on the digital version of the test. Equivalence would allow raw scores to be interchangeable and preexisting norms to be applicable to the digital format of the WAIS-IV. The first digital adaptation continued to use printed response booklets for administration of processing speed subtests to keep formats as similar as possible.

Two studies were conducted in 2011 to assess equivalence between paper-pencil and digital formats of 15 WAIS-IV subtests (Daniel, 2012a). Study 1 used a non-randomly assigned equivalence design to assess equivalence. The WAIS-IV normative sample (i.e., paper-pencil format) used for comparison consisted of 2,200 participants between the ages of 16 and 77. Recruited participants were randomly divided into two substudies: Study 1a and Study 1b, with 39-40 examinees in each substudy. The substudy population included nonclinical samples only (Daniel, 2012a).

Results of Study 1 indicated effect sizes (i.e., the average residual divided by three) of 12 of the 15 subtests were less than 0.2. Therefore, equivalence of Q-interactive administration of the 12 subtests to the standard procedure was demonstrated. Digital Administration of Information, Picture Completion, and Coding exceeded the 0.2 criterion. Scores on Information (effect size = -0.28) and Picture Completion (effect size = -0.43) were lower and scores on Coding (effect size = 0.34) were higher than expected. Scores on digital administration of Symbol Search and Cancellation were also higher than expected, though not significantly (i.e., effect sizes 0.18 and 0.16, respectively, did not exceed 0.2). The aforementioned subtests were subjected to a careful review of video recordings of administration and of potentially invalid data. Picture Completion was updated, but no explanations were discovered for format effects of Information, Coding, Symbol Search, or Cancellation (Daniel, 2012a).

The purpose of Study 2 was to address discrepancies between predicted and actual scores (i.e., an indication of potential format effects) detected in Study 1. Part of Study 2 used a non-randomly equivalent group design (N = 99) to further analyze Information and Picture Completion subtests. For this analysis, the digital format of Picture Completion was updated to include clearer images. The examiner's screen was redesigned to make scoring more seamless. Information was not changed. These decisions were made secondary to review of paper-pencil and digital administrations of each subtest to consider for potential causes of the discrepancy.

Changes made to Picture Completion reduced the effect size from -.43 in Study 1 to -.17 in Study 2. Information remained outside of the range (effect size from Study 2 was -.28 compared to -.29 in Study 1) deemed acceptable to demonstrate equivalence. Additional analysis of Information included applying the same analyses run on digital format (i.e., running a multiple correlation) to Information in paper-pencil format. Data collected in Study 1 from 40 participants was used. The multiple correlation for Information in paper-pencil format was .74 compared to .84 for Information in digital format. The effect size was -.17 compared to -.29 in Study 1. Daniel (2012a) concluded scores were lower than predicted regardless of whether Information was administered in digital or paper-pencil format; it was stipulated that the cause of the effect is unknown.

WISC-IV. Similar to the WAIS-IV, the WISC-IV was adapted to a Q-interactive platform (Daniel, 2012b). A study of equivalence was conducted between the WISC-IV paper-pencil and the WISC-IV digital formats. The goal was to demonstrate raw-score equivalence between paper-pencil and digital formats to apply existing norms from paper-pencil format to digital format. Paper-pencil response booklets continued to be used for administration of processing speed subtests. A randomly equivalent groups design was used for the WISC-IV equivalence study. The sample included 344 children, ages 6 to 16. Each participant was randomly assigned to either the digital format or the paper-pencil format. Researchers required an even division of formats within each age group, gender, SES level, and ethnicity. Each examinee was administered the WISC-IV in her assigned format followed by the Kaufman Brief Intelligence Test, Second Edition (KBIT-2; Kaufman, 2004); the Speed of Information Processing subtest of the Differential Ability Scales, Second Edition (DAS-II; Elliott, 2007); and the Letter Span subtest from the WISC-IV Integrated (Daniel, 2012b) in paper-pencil format.

The paper-pencil tests served as covariates. Collected data was reviewed for quality. Researchers inspected pairs of WISC-IV subtests and pairs of WISC-IV subtests and covariate tests using bivariate scatterplots of scaled scores. Data from two participants were identified as outliers and excluded after researchers reviewed differences between actual scores and predicted scores. Predicted scores were developed using a regression model and were based on demographics and covariate tests.

A multiple regression was run on each WISC-IV subtest in which the dependent variable was the subtest's scaled score and the independent variables were demographics (gender, SES [1-5], ethnicity [dummy codes for groups other than White], covariate tests, and format [0 = paper-pencil; 1=digital]. Format effects were measured by dividing the unstandardized regression weight by 3. Data from the 175 examinees tested with paper-pencil was used to generate prediction equations. The effect of digital format was evaluated using the residual score calculated by subtracting a participant's predicted score from her obtained score. Finally, the relationship of the residual scores to population characteristics was considered by running a linear correlation for continuous variables (ability, age, and SES), a t-test for gender (with unequal variance) and an analysis of variance for ethnicity (coded as African American, Hispanic, White, and Other).

All subtests had strong multiple correlations with demographic variables and covariate tests between .50 and .75 except Cancellation (*r* = .39). Matrix Reasoning and Picture Concepts had effect sizes greater than .20 (effect sizes were .27 and .21, respectively). Researchers performed 75 statistical tests and 7% (five) had statistically significant results. These results were assumed to have occurred by chance and were not interpreted as evidence of a systematic effect as they occurred across multiple variables. Scores from digital administration of Letter Number

Sequencing and Picture Completion were significantly higher than scores from paper-pencil administration. Scores from two subtests correlated negatively with age: younger examinees tended to perform better on Picture Completion (*r* = .16; p < .05) and Word Reasoning (*r* = .20; p < .05). SES had a positive correlation with one subtest: participants performed better on digital administration of Arithmetic (*r* = .17; p < .05) in cases in which parents had higher levels of education. There was a significant effect of gender on two subtests: females performed better on digital administration of Picture Completion (t = 2.14; p < .05) and Picture Concepts (t = 2.00; p < .05).

WISC-V. Daniel et al. (2014) evaluated the equivalence of digital and paper-pencil administration for 18 WISC-V subtests using a randomly equivalent groups design. The sample population included 350 nonclinical participants. Using a randomly equivalent groups design, participants were randomly assigned to the digital condition or the paper-pencil condition. Subsequent to data collection, each case assigned to the digital format was paired with a case assigned to paper-pencil format that matched in age, gender, ethnicity, and parent education (175 total matched pairs). Matched-pairs, along with covariate subtest administration, were intended to increase experimental control and statistical power. Results did not reveal significant statistical differences between subgroups (i.e. age, gender, ethnicity, socioeconomic status, and ability). It was determined examiners had sufficient experience with administration, having participated in seven previous equivalence studies. Therefore, the WISC-V equivalence study administrations were not video recorded for review (Daniel, Wahlstrom, & Zhang, 2014).

Processing speed subtests were not included in this WISC-V equivalence study (Daniel, Wahlstrom, & Zhang, 2014). The initial goal was to demonstrate newly developed WISC-V paper-pencil and digital versions measured the same construct and had similar psychometric properties. This would make it possible to equate paper-pencil and digital versions for use as alternative forms of intelligence measures. The examiner experience with administration on a digital platform remained consistent with WISC-IV Q-interactive experience, and required the examiner to use the tablet from which to read instructions and keep time during administration.

Evaluators reviewed the difference between examinee scores on digital subtests and predicted scores (based on covariate WISC-V subtest scores as well as demographic variables [ability, age, and socioeconomic status]) to assess format effects. The subtests with relatively low format effects were Visual Puzzles, Picture Concepts, Picture Span, Letter-Number Sequencing, Similarities, and Vocabulary; scores on these subtests were used in stage two of the study as predictor variables. Three subtests had a statistically significant format effect: Block Design, Comprehension, and Arithmetic. Effect sizes did not exceed 0.20 (Block Design = 0.20; Comprehension = -0.20; Arithmetic = -.16), though Block Design and Comprehension were at threshold. Generally, results also indicated no differential effects exist for demographic factors, although results were statistically significant for the Naming Speed Quantity subtest. On Naming Speed Quantity in the digital format, older participants and male participants had higher scaled standard scores than younger participants and female participants, respectively. Therefore, it was determined all 18 subtests met standards for equivalence in paper-pencil and digital formats (Daniel, Wahlstrom, & Zhang, 2014).

Processing Speed Equivalence Studies

It was assumed that paper-pencil and digital versions of processing speed measures would not have equivalent raw scores should processing speed subtests be fully administered on the iPad. Therefore, during equivalence studies of the WAIS-IV, WISC-IV, and WISC-V the examinee experience during processing speed subtests remained consistent regardless of format (e.g., the examinee responded with a pencil in a paper response booklet). The only difference between digital administration and paper-pencil administration occurred on the examiner's end. The examiner was required to use an iPad from which to read the directions and time the administration (Daniel, 2012a; Daniel 2012b; Daniel, Wahlstrom, & Zhang, 2014). Therefore, researchers only included processing speed subtests in the WAIS-IV equivalence study, the first Wechsler paper-pencil and digital format equivalence study.

Processing speed subtests from the WAIS-IV equivalence study yielded statistically significant, or nearly statistically significant results in Study 1 (Daniel, 2012a). Study 2 was conducted to further analyze these results. Part of Study 2 included Coding, Symbol Search, and Cancellation (N = 30; 15 matched pairs) and used a retest design. Each participant was administered Coding, Cancellation, and Symbol Search in digital and paper-pencil formats. A matched-pairs t-test was used for analysis; half of participants were administered the digital format first and half were administered the paper-pencil format first. Processing speed subtest scores in digital format were higher than scores in paper-pencil format. This is consistent with results from Study 1. However, developers identified a timing error that allotted 2% more time on processing speed subtests during digital administration. To correct for this imbalance, Daniel (2012a) reduced participants' raw scores by 2%. This reduction changed the effect size from 0.12 to 0.07 for Coding and from 0.27 to 0.13 for Symbol Search. The effect size for Cancellation remained the same. The 2% timing error was immediately corrected and is not relevant for subsequent digital administrations.

The way in which the timing error identified on processing speed measures was corrected is a potential drawback to the WAIS-IV equivalence study. A solution able to be retroactively applied was preferable because data for Symbol Search and Coding had already been collected.

Reducing participants' raw scores by 2% to correct for a 2% timing error assumes the examinee's response rate is consistent throughout the two-minute allotment of time. This does not account for the probability that as each processing speed subtest progresses, participants become more fluent with regard to response rate and speed.

In 2016, two years after the release of the WISC-V, researchers adapted WISC-V processing speed subtests into digital format. In the release of their *Technical Report: WISC-V Coding and Symbol Search in Digital Format: Reliability, Validity, Special Group Studies, and Interpretation,* Raiford and Zhang et al. (2016) highlight substantial changes made to the digital format of the PSI subtests including, "onscreen touch responses, scrolling stimuli, and the elimination of writing requirements and self-corrected responses" (p. 2). An equivalence study was conducted with the WISC-V Processing Speed Index (PSI) subtests, Symbol Search in digital format (SS-D) and paper-pencil format (SS-P), Coding in digital format (CD-D) and paper-pencil format (CD-P), and Cancellation in digital format (CA-D) and paper-pencil format (CA-P). The goal of this study was to "establish a scaling relationship" (a process requiring the conversion of raw scores to scores on the same standard scale) between CD, SS, and CA subtests in paper-pencil and digital formats to assure interchangeability for clinical use (Raiford & Zhang et al., 2016). The study was divided into four research stages: the Conceptual Development Stage, Pilot Stage, Standardization Stage, and Final Assembly and Evaluation Stage.

During the Conceptual Development Stage, the primary goal was to develop a digital format for administration of the PSI subtests that did not require the examinee to use a paper response booklet or a pencil. A number of sources were consulted during this phase to collect information about digitizing the format and maintaining similarity with paper-pencil administration. Sources consulted included user interface designers, cognitive ability testing experts, and literature regarding computerized testing (Raiford & Zhang et al., 2016)

The Pilot Stage consisted of Pilot 1, Minipilot 1 and 2, Pilot 2, and Pilot 3. From the beginning, it was assumed that raw-score equivalence would not be attainable for paper-pencil and digital versions due to difference in response format. Therefore, during Pilot 1, the goal was to demonstrate sufficient correlation between paper-pencil and digital formats and subsequently apply equating procedures. The target during adaptation of the PSI subtests was to make as few changes as possible from paper-pencil to digital format. Consequently, participants were required to "turn the page" by touching an arrow at the bottom of the screen for CD-D and SS-D. For CD-D, participants used a stylus to draw designs. While page-turning was not required for CA-D, the subtest was divided into four display screens and participants were allotted 15 seconds to complete each display screen.

Pilot 1 was completed in conjunction with the WISC-V standardization stage (Daniel, Wahlstrom, & Zhang, 2014) previously described in detail. During Pilot 1, participants were able to self-correct responses (Raiford & Zhang et al., 2016). Results of Pilot 1 revealed issues; equivalence between WISC-V standardization paper-pencil format and Pilot 1 digital format was inadequate for use of equating procedures. From an empirical perspective, not all correlations reached the predetermined threshold (i.e., r = .7). The correlation between CD-D and CD-P was moderate, less than r = .4 for ages 6-7; additional correlations observed were not provided. From a process response perspective, a video review revealed the digital features such as page-turning and initial pause and reorientation to a new page invoked cognitive processes other than processing speed (i.e., selective attention, cognitive flexibility, additional attention; Raiford & Zhang et al., 2016). Additionally, the stylus was deemed problematic. Children tended to have difficulty maintaining their grip on the stylus and would inadvertently obstruct the tablet's registration of responses. For example, children unintentionally touched the tablet screen with the sides of their hands, which led to accidental response captures.

During Minipilot 1, a few alterations to instructions and hardware were made to address the issues identified with process response. Alterations included additional instructions to address the page switching effect and a larger stylus to address problems with response acceptance. A sample of 30 participants, ages 6-7 were administered the PSI subtests in digital format (PSI-D) with the aforementioned alterations. This age range was used because the largest effects were seen in this population. Results from Minipilot 1, in conjunction with consultation with experts, indicated designs with smoother transitions (e.g., scrolling stimuli instead of page-turning) were necessary. Elimination of a stylus was also indicated. Finally, it was concluded CA-D could not be successfully adapted into a digital format and therefore continues to be administered with a response booklet (Raiford & Zhang et al., 2016).

Three prototypes of CD-D and three prototypes of SS-D were created for Minipilot 2 based on feedback gathered in Minipilot 1. The six prototypes were submitted for expert review prior to being administered to a convenience sample (N = 10). An advisory panel of experts was then assembled to examine administration and prototypes. Surveys and interviews were used to collect data on user friendliness, examinee behaviors, response processes, and construct measured (Raiford & Zhang et al., 2016). It was concluded the prototypes were acceptable to use for a larger pilot study.

For Pilot 2, a single SS-D subtest and a single CD-D subtest were created by combining aspects of the three prototypes for each subtest used in Minipilot 2. The goals of Pilot 2 were to demonstrate adequate correlation between paper-pencil and digital formats of CD and SS (i.e., greater than *r* = .70), demonstrate strong correlations between digital versions of CD-D and SS-D (i.e., greater than *r* = .50), and review response processes. A sample of 100 participants (n = 50 ages 8-9; n = 50 ages 6-7) was used. Researchers reported that correlations exceeded the set thresholds, though exact correlations were not included.

There were three main goals of Pilot 3: (a) review the distribution of scores for CD-D and SS-D, (b) examine the intercorrelations of CD-D and CD-P and of SS-D and SS-P, and (c) examine correlations between CD-D and SS-D and the additional eight primary subtests. Processing speed subtests were placed in the Q-interactive platform and participants were randomly assigned to digital format administration or a waitlist condition. Participants assigned to the waitlist condition did not participate in Pilot 3 and instead participated in the Standardized Stage. Pilot 3 featured a sample size of 70 participants (n = 35 ages 6-7; n = 35 ages 8-9) that were noted to "match census proportions within 4% of target" (Raiford & Zhang et al., 2016, p.5). Researchers concluded CD-D and SS-D "related to each other as expected" (Raiford & Zhang, et al., 2016, p.5). No additional reports were included concerning intercorrelations or distributions of scores. No changes were made to subtests. During Pilot 3, samples from populations with intellectual giftedness, intellectual disabilities, attention-deficit/hyperactivity disorders, and autism-spectrum disorders were tested to assess interface and reactions to digital test administration. Raiford and Zhang et al. (2016) indicated special populations did not appear to experience adverse reactions to digital stimuli.

The goals of the Standardization Stage included the development of standard scores and gathering of reliability, validity, and clinical utility evidence of CD-D and SS-D. This Stage used a stratified sample of 329 participants (45 and 49 participants for the first two age groups, respectively and 20-30 participants for each of the remaining age groups: 6.0-6.11, 7.0-7.11, 8.0-

8.11, 9.0-9.11, 10.0-10.11, 11.0-11.11, 12.0-12.11, 13.0-13.11, 14.0-14.11, 15.0-15.11, and 16.0-16.11). For context, during standardization of the WISC-V, data from a normative sample of 2,200 participants was collected; each of the 11 age groups included 200 participants (Wechsler, 2014). During the Standardization Stage of the processing speed measures, participants were administered the WISC-V standard battery in digital format, including CD-D and SS-D, in order (i.e., Block Design, Similarities, Matrix Reasoning, Digit Span, CD-D, Vocabulary, Figure Weights, Visual Puzzles, Picture Span, then SS-D). They were subsequently administered CD-P and SS-P. It was determined that standardization goals were met and evidence of reliability and validity was demonstrated. The raw score correlation was r = .87 for CD and r = .84 for SS. The scaled score correlation was r = .63 for CD and r = .67 for SS. The standard difference was .23 for CD and -0.13 for SS. Researchers recognized that .23 effect size for CD exceeded the standard for equivalence. The correlation for the PSI in digital format and paper-pencil format was not provided.

In the Final Assembly and Evaluation Stage, the CD-D and SS-D subtests remained the same. A test-retest coefficient was chosen over the split-half coefficient. Stability coefficients for the overall sample were .80 for CD and .78 for SS and had a standard difference (i.e., difference between the mean score of the first testing and mean score of the second testing divided by the square root of the pooled variance) of .29 and .54, respectively. To establish a relationship by converting raw scores of paper-pencil and digital versions of CD and SS to scores on the same standard scale, the inferential scaling method (Zhu & Chen, 2011) was used. This allowed scores from paper-pencil and digital versions to be interpreted on the same metric, as scaled scores. Scaled scores were created for digital administration by applying the inferential scaling method to digital raw scores. Subsequent to scaling procedures, analyses determined that digital and paper-pencil scaled scores demonstrated no meaningful differences.

Research Questions and Predictions

The aim of the current study was to assess whether the paper-pencil and digital versions of the WISC-V Processing Speed Index and comprised subtests (CD and SS) are interchangeable. The current study answered three primary research questions:

1. How highly are standard scores correlated on (a) CD-D and CD-P, (b) SS-D and SS-P, and (c) the PSI-D and the PSI-P? To analyze the paper-pencil and digital format equivalence, a concurrent validity design was used. Pearson correlations were calculated between scaled scores yielded by paper-pencil and digital formats of SS, CD, and the PSI. It was hypothesized correlations between scores on paper-pencil and digital formats of processing speed subtests would be less than .79 (the correlation between the WISC-V PSI-P and the WAIS-IV PSI-P; Wechsler, 2014) and .70 (the standard for equivalence set by Pearson in similar studies; Daniel, 2012a; Daniel, 2012b). It was also hypothesized that subtests designed to measure the same construct would correlate higher than subtests sharing the same method (e.g., the correlation between SS-D and SS-P [that share the same construct] would exceed the correlation between SS-D and CD-D [that share the same method]).
2. Are means of standard scores nonequivalent on (a) CD-D in comparison to CD-P, (b) SS-D in comparison to SS-P, and (c) the PSI-D in comparison to the PSI-P? A two-tailed paired t-test was conducted to determine whether scores from the digital format differed from scores from the paper-pencil format for CD, SS, and the PSI. It was hypothesized that scaled scores on digital measures would not be equivalent to scores on paper-pencil measures.

3. Is visual-motor integration as indicated by the Beery-Buktenica Developmental Test of Visual-Motor Integration (Beery VMI; Beery & Beery, 2010) more related to performance on the paper-pencil format of the processing speed tests compared to the digital format. Difference scores were calculated by subtracting the standard score of the PSI-P from the standard score of the PSI-D, the scaled score of CD-P from the scaled score of CD-D, and the scaled score of SS-P from the scaled score of SS-D, respectively. Pearson correlations were calculated between the PSI-Difference (PSI-Diff), Coding-Difference (CD-Diff), and Symbol Search-Difference (SS-Diff) scores; and the aforementioned Beery VMI scores. It was hypothesized that scores on the Berry VMI would negatively correlate with all difference scores, demonstrating the impact of psychomotor skills on CD-P and SS-P performance. Also, it was hypothesized scores on the Beery VMI would more positively correlate with scores on the paper-pencil versions of the PSI, CD, and SS than on digital versions of the PSI, CD, and SS.

Chapter II

Method

Participants

The sample included 41 participants between the ages of 13 years old and 16 years and 11 months old. Participants were recruited from a private school in New Jersey. All measures were administered to participants individually in the school's library. Participants in the sample were predominantly European American (85.3%) and received general education services (85.4%). About half of participants identified as female (53.7%) and half as male (46.3%). The sample varied with regard to the highest degree attained. The predominant subgroup included participants that reported years of education completed by mother as 16 years or more (75.6%). The predominant subgroup included participants that reported years of education completed by father as 16 years or more (61%). Table 7 further depicts the demographic characteristics of the students in the sample.

Measures

A demographic questionnaire was used to gather information on the sample. Two processing-speed subtests from a test of cognitive functioning were administered using Q-interactive in both paper-pencil and digital versions to gather data on discrepancies in performance based on format. Additionally, an assessment of visual-motor functioning was administered to assess for psychomotor difficulties.

Table 7.

Participant Demographics

Characteristic	*n*	%
Gender		
Female	22	53.7%
Male	19	46.3%
Age		
Ranges		
13-13.11	2	4.9%
14-14.11	15	36.6%
15-15.11	8	19.5%
16-16.11	16	39%
Parental Education: Mother		
11 years or less	0	0%
12 years	7	17.1%
13-15 years	2	4.9%
16 years or more	31	75.6%
Unknown	1	2.4%
Parental Education: Father		
11 years or less	0	0%
12 years	10	24.4%
13-15 years	3	7.3%
16 years or more	25	61.0%
Unknown	3	7.3%
Ethnicity/Race		
African American	4	9.7%
Asian	1	2.4%
Hispanic or Latino/a	1	2.4%
European American	35	85.3%
Disability Classification		
General Education	35	85.4%
General Education + 504	2	4.9%
Special Education	4	9.8%
Total	41	

Wechsler Intelligence Scale for Children – Fifth Edition. The WISC-V was released in 2014 and is used to assess the cognitive functioning of children and adolescents between the ages of 6 years old and 16 years and 11 months old (Wechsler, 2014; Wechsler, 2016). WISC-V results include subtest and composite scores that represent intellectual functioning in specific cognitive domains. Composite scores are presented as standard scores and have an average of

100 and a standard deviation of 15. Subtest scores are presented as scaled scores and have an average of 10 and a standard deviation of 3. Standard and scaled scores were obtained using age based norms.

The composite of interest is the Processing Speed Index (PSI). There are three subtests that are categorized as Processing Speed Subtests: Coding (CD), Symbol Search (SS), and Cancellation (CA). The processing speed subtest CA was not included in the current study. The reasoning is two-fold. First, it was not able to be adapted to a digital format and is therefore only administered with a response booklet (Raiford & Zhang et al., 2016). Second, it is a supplemental subtest and is not required to derive the PSI or Full Scale Intelligence Quotient (FSIQ) scores. As it stands, users of Q-interactive are offered a choice to either administer the digital or the paper-pencil version of CD and SS during iPad administration of the WISC-V.

WISC-V PSI: Paper-pencil version. The PSI paper-pencil format (PSI-P) measures speed and accuracy of visual identification, decision-making, and decision implementation. Performance on the PSI-P requires visual scanning, visual discrimination, short-term visual memory, visual-motor coordination, and concentration (Wechsler, 2016). The PSI-P is comprised of two subtests: Coding (CD-P) and Symbol Search (SS-P). At the time of release in 2014, the WISC-V PSI subtests were only available in paper-pencil format.

Normative information for the WISC-V was collected from a sample of 2,200 children. Demographic characteristics including race/ethnicity, age, sex, parent education level, and geographic region were obtained and closely resembled the data from the 2012 U.S. census proportions. Ethnic groups sampled included: White, African American, Hispanic American, Asian American, and Other. The geographic regions sampled included: Northeast, South, Midwest, and West. The parent levels of education sampled included: Eight years or less, Some high school, High school graduate, Some college, and College graduate (Wechsler, 2014). Participants between the ages of 6 years-old and 16 years- and 11 months-old were divided into eleven age groups. Each of the eleven age groups included 100 boys and 100 girls (Wechsler, 2014).

Test-retest reliability is a measure of how stable responses are whenever a respondent completes a measure more than once (Litwin, 1995, p.8). Developers assessed test-retest reliability by retesting participants following an interval between 9 and 82 days, with a mean of 26 days. Participants included approximately 220 children and were divided into five age groups. The average test-retest reliability estimate for the PSI was .81 (range = .77-.88), indicating that the PSI scores are generally stable (Sattler, Dumont & Coalson, 2016). The test-retest reliability coefficient for the PSI was .88 (range = .84-.92). The average Standard Error of Measurement (SEM) for the PSI in standard-score points was 5.24 (range = 4.24-6.00). The intercorrelation between the two subtests that contribute to the PSI (CD-P and SS-P) was .58; this shared correlation is the lowest compared to other pairs of scores from a common index (Wechsler, 2014). CD (*r* = .50) and SS (*r* = .46) had the lowest correlations of the ten core subtests (range = .46 - .77) with the Full Scale Intelligence Quotient (FSIQ; Wechsler, 2014).

Criterion Validity was assessed using the PSI correlations between the WISC-V and four other measures of intelligence: Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV), Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI-IV), Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV), and Kaufman Assessment Battery for Children-Second Edition (KABC-II). Pearson correlation was *r* = .70 with the WISC-IV PSI, *r* = .34 with the WPPSI PSI, *r* = .79 with the WAIS-IV PSI, and *r* = .04 with the KTEA-3 Mental Processing.

Coding. On CD-P, the examinee is required to copy simple geometric shapes into empty boxes using a key. Participants between ages 9 and 16 are administered Form B, which includes 117 test items. A scoring template is used to score all items. Raw scores are obtained by calculating the sum of each item correctly answered. Rotation errors are recorded. A rotation error is identified as a correctly drawn symbol that is rotated 90, 180, or 270 degrees. Rotation errors yield a process score calculated using age-based norms to help determine the nature of errors committed by the examinee.

Task description. To complete the task, the examinee is provided a pencil without an eraser and the WISC-V response booklet. The examiner opens to page five and places the response booklet in front of the examinee. A key is located on the top of the page with nine separate boxes. The boxes are divided in half: the top half of each box includes a number (1-9) and the bottom half of each box includes a unique geometric shape. Below the key are seven rows with 18 boxes in each row. The top half of each box has a number (1-9) and the bottom half of each box is empty. The examinee is asked to draw the shape that is paired with the given number in the empty box using the aforementioned key as a reference. The examiner demonstrates the task by completing the first three items for the examinee while providing commentary. Next, the examinee is asked to complete six items as a sample administration. The examiner immediately corrects and explains errors made on sample items. After administration of sample items, the examinee is asked to complete as many items as possible until she has reached the last item or until the examiner says “stop.” One hundred and twenty seconds is allotted for the task.

Psychometrics. The average test-retest reliability estimate for CD-P was .79 (range = .73 -.85), indicating that CD-P scores are generally stable (Sattler, Dumont & Coalson, 2016). A t-test used a repeated-measures formula to evaluate the mean change on each subtest. Test-retest intervals had a mean of 26 days and ranged from 9 to 82 days. The mean change was 1.3 for the total sample, which was significant at the p < .001 level with a small effect size of .43. The sample used to assess test-retest reliability, approximately 220 children, was divided into five age groups. The age groups relevant to the study include: twelve to thirteen (0.7) and fourteen to sixteen (1.1).

The average internal consistency reliability coefficient (test-retest reliability coefficient) for CD-P was .82 (range = .78 - .86). The average Standard Error of Measurement (SEM) for CD-P in standard-score points was 1.28 (range = 1.12 - 1.41). The correlation between CD-P and the PSI was *r* = .89. CD-P had the second lowest correlation with the FSIQ (*r* = .50) of the ten subtests included in the core battery. CD-P correlations with the other core subtests were small, as hypothesized (Wechsler, 2014). Researchers used oblimin rotation, specifying five factors unique from results in the Technical and Interpretative Manual (Wechsler, 2014). Sattler, Dumont, and Coalson (2016) conducted an exploratory principal axis factor analysis on the ten core WISC-V Subtests. For the PSI-P, the factor is the ability to “process visually perceived nonverbal information quickly, with concentration and rapid eye-hand coordination being important components” (Sattler, Dumont & Coalson, 2016). CD-P loadings on the processing speed factor were calculated for eleven age groups. Relevant age groups include: thirteen (.83), fourteen (.45), fifteen (.88), and sixteen (.86). The total CD-P loading on the processing speed factor is high (.87).

Symbol Search. On SS-P, the examinee is required to scan a search group and indicate whether a target symbol matches any of the symbols in the search group within a specified time limit. Participants between the ages of 8 and 16 are administered Form B, which includes 60 test items. A scoring template is used to score each item. Raw scores are calculated by subtracting the number of incorrectly answered items from the number of correctly answered items. Two special types of errors are possible: set errors and rotation errors. For a response to be considered a set error, the examinee must incorrectly mark a symbol that is perceptually similar to one of the target symbols. For a response to be considered a rotation error, the examinee must mark an item identical in form to a target symbol that is rotated 90, 180, or 270 degrees. A process score can be yielded using age based norms for both set and rotation errors.

Task description. To complete the task, the examinee is provided a pencil without an eraser and the WISC-V response booklet. The examiner opens to page 13 and places the response booklet in front of the examinee. Page 13 includes two demonstration items and three sample items. Each item is a single row that includes two target symbols, five symbols to be scanned, and a box that includes the word "no." The examiner explains the task to the examinee by completing the demonstration items while providing commentary. Next, the examinee is asked to complete the three sample items. The examiner immediately corrects any errors using a correction script. Upon completion of the sample items, the examiner turns the page to reveal the first two of six pages of test items. Each page includes 10 rows (i.e. ten items per page). The examinee is allotted 120 seconds to complete as many items as possible, in order. To complete an item, the examinee is required to look closely at the two target symbols. Each row has its own unique set of target symbols. Then, the examinee scans five symbols to determine whether a symbol in the row matches a target symbol exactly. If so, the examinee is asked to draw a line through the symbol (i.e. a slash mark). If not, the examinee is asked to draw a line through the box containing the word "no."

Psychometrics. The average test-retest reliability estimate for SS-P was .76 (range = .62 - .84), indicating that SS-P scores are generally stable (Sattler, Dumont & Coalson, 2016). The mean change was 1.5 for the total sample (220 children), which is significant at the p<.001 level with a medium effect size of .51. The mean change for relevant age groups was: twelve to thirteen (1.7) and fourteen to sixteen (1.3).

The average internal consistency reliability coefficient (test-retest reliability coefficient) for SS-P was .81 (range = .67 - .87). The average Standard Error of Measurement (SEM) for SS-P in standard-score points was 1.34 (range = 1.08 - 1.72). The correlation between SS-P and the PSI was *r* = .89. SS-P had the lowest correlation with the FSIQ (*r* = .46) of the ten subtests in the WISC-V core battery. SS-P correlations with all other subtests were small, as hypothesized (Wechsler, 2014). SS-P loadings on the processing speed factor were calculated for eleven age groups. Relevant groups include: thirteen (.75), fourteen (.60), fifteen (.79), and sixteen (.82). The total SS-P loading on the processing speed factor is high (.82).

WISC-V PSI: Digital version. In 2016, Pearson released the PSI subtests, CD and SS, in digital format (Raiford & Zhang et al., 2016). During administration of CD-D and SS-D two iPads are used. The examiner's iPad screen is kept in an upright position, out of sight of the examinee. The examinee's iPad is laid flat on the table, facing the examinee. A digital platform, *Q-interactive Assess*, is utilized to connect the iPads via Bluetooth and securely administer WISC-V digital subtests. In addition to instructions and prompts, the examiner's iPad screen displays the information the examinee is seeing throughout administration. In cases in which an examinee incorrectly answers an item, a transparent red circle appears over the examinee's item choice on the examiner's screen. In cases in which an examinee correctly answers an item, a transparent, green circle appears.

A scaling relationship between the paper-pencil and digital formats of WISC-V subtests was established using the inferential scaling method in order to assure practitioners either format is acceptable for clinical use (Raiford & Zhang et al., 2016). To establish a relationship using the inferential scaling method, raw scores from each format were transformed into scores on the same standard scale. Principal differences between the paper-pencil and digital formats of the PSI subtests (i.e. touch responses, presentation of stimuli, elimination of written responses, and elimination of self-corrections) were found. Therefore, prior to establishing a scaling relationship, evidence of reliability and evidence of validity were first derived from newly collected data on CD-D and SS-D to determine whether digital and paper-pencil formats have similar psychometric properties and measure the same construct (Raiford & Zhang et al., 2016).

Coding. On CD-D, the examinee is asked to choose a geometric shape that corresponds with a number using a key. Participants between ages 9 and 16 are administered Form B, which includes 117 test items. The Q-interactive system processes scores automatically and includes data on duration between each response and whether the response was correct. Rotation errors are not applicable for CD-D responses; therefore, no process scores are calculated.

Task description. Throughout administration of CD-D, the examinee's screen displays one item at a time. Each item includes three parts. The first part is the answer key. The answer key is located at the top of the screen and includes nine boxes. Each box is divided in half: the top half of each box includes a number (1-9) and the bottom half of each box includes a unique geometric shape. The second part, directly below the answer key, is a single box divided in half with a number in the top half of the box and an empty space in the bottom half of the box (i.e. the number-only box). The third part, directly below the number-only box, is a row of five geometric shapes. Included within this row is the geometric shape corresponding to the number in the number-only box (i.e. the correct response). The examinee is required to respond by tapping on the geometric shape corresponding with the number in the number-only box.

The examiner begins the subtest by tapping the ‘start’ button on the examiner’s screen. The first demonstration item appears on the examinee’s screen. The examiner then explains the subtest directions and completes three demonstration items, one by one, by tapping the correct response on the examinee’s screen. After the demonstration items, the examinee’s screen appears grey. The examiner is then required to click continue on the examiner’s screen. Subsequently, the first sample item appears on the examinee’s screen. During sample items, if an examinee makes an error, a red circle appears on the examiner’s screen. The examiner is then required to read a correction script before proceeding to the next sample item. Administration of sample items will not advance until the examinee has selected a correct response. After administration of six sample items, the examinee’s screen again turns grey and the examiner reads further test instructions. Once the examinee does not have additional questions, and understands directions, the examiner administers the test items. The examinee is allotted 120 seconds to complete as many items as possible. After an examinee chooses a response, the next item appears on her screen automatically, regardless of whether the answer is correct or incorrect. No self-corrections are allowed during test item administration. At the end of 120 seconds, a check mark appears on the examinee’s screen, indicating the end of the subtest.

Psychometrics. The average test-retest stability coefficient for CD-D for ages eight to sixteen was .85. The effect size (i.e., difference between the mean scores of the first and second testing divided by the square root of the pooled variance; [M_1-M_2] / SD) between the first and second testing for all ages was .34. Test-retest intervals had a mean of 24.8 days and ranged from 14 to 52 days. The sample used to assess test-retest reliability was approximately 35 children between the ages of eight to sixteen.

The correlation between CD-D and the PSI was *r* = .88. CD-D had the second lowest correlation of the ten core subtests with the FSIQ (*r* = 45). CD-D correlations with all other subtests were small, as hypothesized (Raiford & Zhang et al., 2016). A confirmatory factor analysis was conducted on the overall age range (six to sixteen). The total CD-D loading on the g factor (overall intelligence) is .32. Correlations between CD-P and CD-D were calculated to assess format equivalence. The raw score correlation was *r* = .87. The scaled score correlation was *r* = .63. The standard difference was .23. Researchers recognized an effect size of .23 for CD exceeded the standard for equivalence (set at .20).

Symbol Search. On SS-D, the examinee is required to scan a search group and indicate whether one of the two target symbols matches a symbol in the search group within a specified time limit. Participants between the ages of 8 and 16 are administered Form B, which includes 60 test items. The Q-interactive system processes scores automatically and includes data on duration between each response and error scores. Two types of errors are possible: set errors and rotation errors. For a response to be considered a set error, the examinee must incorrectly mark a symbol that is perceptually similar to one of the target symbols. For a response to be considered a rotation error, the examinee must mark an item identical in form to a target symbol that is rotated 90, 180, or 270 degrees. Q-interactive calculates process scores for both set and rotation errors.

Task description. During administration, three rows appear on the screen; each row is a separate item. The middle row is considered "active" and is displayed in black font. The rows above and below the active row are displayed in grey font as to appear faded. The iPad only registers participants' responses in cases in which the examinee taps an item in the active row.

After the examinee selects a response, the active row scrolls upward, the bottom row is centered and appears active, and a new row appears faded at the bottom of the screen. Each row includes two target symbols, a search group (composed of five symbols), and a box with the word “no” inside. The examinee is required to respond by tapping either a symbol in the search group or the “no” box.

The examiner begins the subtest by tapping the start button on the examiner’s screen. Two demonstration items will then appear on the examinee’s screen, and one item is active. The examiner then explains the subtest directions and completes the demonstration items, one at a time, by tapping the correct response on the examinee’s screen. After the demonstration items, the examinee’s screen appears grey. The examiner is then required to click continue on the examiner’s screen. Subsequently, the first sample item appears on the examinee’s screen. During sample items, if an examinee makes an error, a red circle appears on the examiner’s screen. The examiner is then required to read a correction script before proceeding to the next sample item. Administration of sample items does not advance until the examinee has selected a correct response. After administration of three sample items, the examinee’s screen again appears grey and the examiner reads further test instructions. If the examinee does not have additional questions, and understands directions, the examiner administers the test items. The examinee is allotted 120 seconds to complete as many items as possible. After an examinee chooses a response, the item is scrolled upward and a new item appears active on the screen. This occurs automatically, whether the answer is correct or incorrect. No self-corrections are allowed during test item administration. At the end of 120 seconds, a check mark appears on the examinee’s screen, indicating the end of the subtest.

Psychometrics. The average test-retest stability coefficient for SS-D for ages eight to sixteen was .78. The effect size (i.e., difference between the mean score of the first and second testing divided by the square root of the pooled variance) between the first and second testing for was .36. Test-retest intervals had a mean of 24.8 days and ranged from 14 to 52 days. The sample used to assess test-retest reliability was approximately 35 children between the ages of eight to sixteen.

The correlation between SS-D and the PSI was *r* = .87. SS-D had the lowest correlation with the FSIQ among the ten core subtests (*r* =.38). SS-D correlations with all other subtests were small, as hypothesized (Raiford & Zhang et al., 2016). A confirmatory factor analysis was conducted on the overall age range (six to sixteen). The total SS-D loading on the g factor (overall intelligence) is .39. Correlations between SS-P and SS-D were calculated to assess format equivalence. The raw score correlation was *r* = .84. The scaled score correlation was *r* = .67. The standard difference was -0.13.

Beery-Buktenica Developmental Test of Visual-Motor Integration – Sixth Edition. The Beery VMI is a norm-referenced measure intended for use with people ages 2-100 to assess "integrating or coordinating visual perceptual and motor (finger and hand movement) abilities" (Beery & Beery, 2010). Between 1964 and 2010, the Beery VMI was nationally standardized six times with more than 13,000 children and 1,021 adults.

Beery-VMI Full Form. The assessment is individually administered and requires the participant to copy increasingly complex two-dimensional geometric forms in a response booklet. For participants between the ages of 8 and 18, the Beery-VMI Full Form includes 30 figures and takes approximately 10-15 minutes to administer. Raw scores are calculated by subtracting the number of attempted items that received no points from the last item administered. Similar to the WISC-V composite scores, the Beery-VMI yields standard scores based on age norms with a mean of 100 and a standard deviation of 15 (Beery & Beery, 2010).

Task description. Participants are asked to use a pencil to copy images and refrain from erasing or rotating the response booklet. Participants at or over the functional age of 5 begin with item 7. Whenever the participant does not earn at least one point on each of the first three items, items 1-6 are administered in order, beginning with item 1. Whenever the participant earns no points on three consecutive items, testing is discontinued. A participant can earn up to one point for each imitated or copied item until testing is discontinued. Any item before the first three correctly administered items receives full credit. A score of one or zero is assigned based on individual item scoring criteria included in the Beery VMI manual. For example, full credit for a circle will be assigned to, "Any loop with a ratio of no more than 2 to 1 between its height and width" (Beery & Beery, 2010, p. 34).

Psychometrics. Evidence of reliability and validity for the Beery VMI is well established (Beery & Beery, 2010). Evidence for internal consistency was established using the Rasch-Wright analysis method. Split-half correlations with one-year age groups had a median of .84 and a high of .93. The test-retest coefficient, based on an average of 14 days between administrations, was .88. Interscorer reliabilities, normed on a variety of professionals, were between .90 and .98. Most recently, concurrent validity was evaluated by comparing the VMI results with the Copying subtest of the Developmental Test of Visual Perception (DTVP-2) and the Drawing subtest of the Wide Range Assessment of Visual Motor Abilities (WRAVMA). Concurrent validity was shown to be moderately high to high. The Pearson correlation between the Beery VMI scores and chronological age was r = .89 (age range: 1 year-old to 95 years-old). As expected, using the Revised Wechsler Intelligence Scale for Children (WISC-R) as a measure of intelligence, scores on the Beery VMI had a strong correlation with nonverbal test results (r = .56) and a moderate correlation with verbal test results (r = .49; age range: 6 years-old through 11 years-old). Among forth and fifth grade students, the Beery VMI correlated higher than expected (r = .63; it was hypothesized the Beery VMI would correlate moderately well with measures of academic achievement) with the Comprehensive Test of Basic Skills (CTBS), a measure of academic achievement. The Beery VMI was also shown to be sensitive to a number of disabilities (Beery & Beery, 2010, pp. 118-120).

Procedures

Doctoral level students from the Graduate School of Applied and Professional Psychology at Rutgers University-New Brunswick administered all assessment measures to participants. Prior to administration, administrators were required to participate in a training conducted by the lead author. Each administrator was exposed to the theoretical background and administration procedures of the measures. The lead author provided assistance to the Doctoral level students throughout administration of measures to participants, as needed.

Each participant completed a general questionnaire. Information gathered included age, gender, race, highest level of education completed by parents, and disability classifications. Subsequently, each participant was administered the CD-P, CD-D, SS-P, and SS-D subtests from the WISC-V. Balanced Latin Square, a counterbalanced measures design, was used to guard against order effects (Bradley, 1958). In experiments with an even number of conditions, the Latin Square formula for the first participant (i.e., row one) is 1, 2, n, 3. In this formula, n = 4 (the number of conditions in this experiment). For subsequent rows, one was added to the previous number, returning to 1 after n. The sequence was repeated for subsequent participants, as needed. This format ensures every condition (1-4) follows every other condition one time, reducing order effects. For the purposes of this study, the Balanced Latin Square was coded: 1 = CD-P, 2 = CD-D, 3 = SS-P, 4 = SS-D (See Table 8).

Table 8.

Counterbalanced Measures Design: Balanced Latin Square

Sequence	**Subtest Ordering**			
A	CD-P	CD-D	SS-D	SS-P
B	CD-D	SS-P	CD-P	SS-D
C	SS-P	SS-D	CD-D	CD-P
D	SS-D	CD-P	SS-P	CD-D

Finally, each participant was administered the Beery VMI to screen for visual-motor deficits. The lead author scored paper-pencil assessments.

Data Analysis

Data analyses were conducted using SPSS Statistics version 24.0. To analyze the paper-pencil and digital format equivalence, a concurrent validity design was used. Pearson correlations were calculated between scaled scores yielded by paper-pencil and digital formats of SS, CD, and the PSI. Fishers r to z transformation was used to assess for a significant difference between the calculated the PSI correlation and .79 (the correlation in previous research between the WISC-V PSI-P and the WAIS-IV PSI-P; Wechsler, 2014). The Multitrait Multimethod Matrix (MTMM) was used to examine evidence for convergent and discriminant validity of measures (Campbell & Fiske, 1959). It was hypothesized the correlations between subtests targeting the same construct (i.e., CD-D and CD-P; SS-D and SS-P) would have higher correlations than subtests using the same format (i.e., CD-D and SS-D; SS-D and SS-P). A two-tailed paired t-test was conducted to determine whether mean scores on the digital format were equivalent to mean scores on the paper-pencil format for CD, SS, and the PSI. The standard difference was calculated by subtracting the mean of the first administered format from the mean of the second administered format. Finally, the difference was divided by the standard deviation to determine effect size, Cohen's *d* (Cohen, 1992; Meyers et al, 2013). Difference scores were calculated by subtracting the PSI-P from the PSI-D, CD-P from CD-D, and SS-P from SS-D yielding scores coded: the PSI-Diff, CD-Diff, and SS-Diff, respectively. Pearson correlations were calculated between the PSI-Diff, CD-Diff, and SS-Diff scores and the aforementioned and the Beery VMI scores. Scores on the Berry VMI were expected to negatively correlate with all difference scores. Pearson correlations were also calculated between scores on the Beery VMI and scores on paper-pencil and digital versions of the PSI, CD, and SS. Scores on the Beery VMI were expected to positively correlate with scores on paper-pencil versions of the PSI, CD, and SS. The correlations were expected to be significantly higher than correlations between scores on the Beery VMI and scores on digital versions of the PSI, CD, and SS, respectively.

Chapter III

Results

Descriptive Data

Descriptive statistics indicated that all variables were within acceptable limits of skewness and kurtosis. The PSI scores measure between one third and one standard deviation above the mean; this indicates the sample population performed higher on processing speed measures than larger population samples. Beery VMI scores measure about one third of a standard deviation below the mean. Table 9 outlines the means and standard deviations for all variables.

Table 9.

Means and Standard Deviations of Scores

	Variable	Mean	Standard Deviation
Standard Scores		**100**	**15**
	PSI-P	108.85	14.15
	PSI-D	110.27	17.94
	Beery VMI	94.98	7.75
Scaled Scores		**10**	**3**
	SS-P	11.83	3.16
	CD-D	11.49	3.33
	SS-D	12.15	3.37
	CD-P	11.15	2.72

Question 1: Correlations among Digital and Paper-Pencil Scores

Pearson correlations were calculated between scores yielded by paper-pencil and digital formats of SS, CD, and the PSI. Pearson correlations indicated that there were significant positive associations between scores on the PSI-P and the PSI-D, $r(39) = .68$, $p < .05$, 95% CI [.49, .82], between scores on SS-P and SS-D $r(39) = .61$, $p < .05$, 95% CI [.39, .77], and between scores on CD-P and CD-D *r*(39) = .67, *p* < .05, 95% CI [.45, .81]. Correlations do not meet the standards set for equivalence (i.e., .70 for SS and CD; .79 for the PSI). There is error in all measurement; the upper limits of the confidence intervals do exceed standards for equivalence. The difference between the correlation of scores on the WISC-V PSI-D and the WISC-V PSI-P (*r* = .68) and the correlation of scores on the WISC-V PSI-P and the WAIS-IV PSI-P (*r* = .79) was not significant (*z* = -1.29, *p* = 0.10, one-tailed).

Table 10.

Correlation Matrix

			PSI-P	CD-P	SS-P	PSI-D	CD-D	SS-D
PSI-P	**Pearson Correlation**		**1**	**.84**	**.87**	**.68**	**.66**	**.59**
	95% CI	Lower	1	.75	.78	.49	.45	.34
		Upper	1	.89	.93	.82	.81	.78
CD-P	**Pearson Correlation**			**1**	**.48**	**.57**	**.67**	**.37**
	95% CI	Lower		1	.31	.33	.45	.09
		Upper		1	.64	.75	.81	.63
SS-P	**Pearson Correlation**				**1**	**.61**	**.49**	**.61**
	95% CI	Lower			1	.41	.27	.39
		Upper			1	.76	.67	.77
PSI-D	**Pearson Correlation**					**1**	**.91**	**.91**
	95% CI	Lower				1	.86	.84
		Upper				1	.96	.95
CD-D	**Pearson Correlation**						**1**	**.67**
	95% CI	Lower					1	.51
		Upper					1	.79
SS-D	**Pearson Correlation**							**1**
	95% CI	Lower						1
		Upper						1

Note: CI = Confidence interval

It was hypothesized subtests designed to measure the same construct would correlate higher than subtests sharing the same method. The correlation between scores on CD-D and CD-P (*r* = .67) exceeded the correlation between scores on CD-P and SS-P (*r* = .48) and was equivalent to the correlation between scores on CD-D and SS-D (*r* = .67). The correlation between scores on SS-D and SS-P (*r* = .61) exceeded the correlation between scores on CD-P and SS-P (r = .48) and was lower than the correlation between scores on CD-D and SS-D (r = .67). The correlation between scores on CD-D and SS-D was higher than expected. The correlation between scores on SS-D and SS-P was lower than expected. Table 10 depicts the correlation matrix between paper-pencil and digital versions of processing speed subtest and index scores.

Question 2: Differences between Mean Scores on Digital and Paper-Pencil Formats

Results of the dependent (paired) samples t-tests indicated there were not significant differences between scores on the PSI-P and the PSI-D $t(40) = -.68$, $p = .497$, $d = .087$, between scores on SS-P and SS-D $t(40) = -.70$, $p = .487$, $d = .096$, or between scores on CD-P and CD-D $t(40) = -.86$, $p = .390$, $d = .11$. Small effect sizes underscore the lack of meaningful difference found between pairs of variables (i.e., the PSI-P and the PSI-D; SS-P and SS-D; CD-P and CD-D). It was hypothesized mean scores on paper-pencil and digital versions of processing speed subtests (CD-P and CD-D; SS-P and SS-D) and index (the PSI-P and the PSI-D) scores would be nonequivalent; this test failed to reject the null hypothesis. Table 11 depicts the mean comparisons.

Table 11.

Mean Comparisons

	Mean	*t*	df	SD	Cohen's d
PSI-P & PSI-D	-1.41	-.69	40	12.22	.087
SS-P & SS-D	-.32	-.70	40	2.89	.096
CD-P & CD-D	-.34	-.87	40	2.52	.11

Question 3: Correlations between Beery VMI and Processing Speed Scores

Results of the Pearson correlations indicated that there were not significant associations between the VMI and the PSI-Diff scores $r(39) = -.15$, $p = .17$, between the Beery VMI and SS-

Diff scores *r*(39) = -.12, *p* = .23, or between the Beery VMI and CD-Diff scores *r*(39) = -.16, *p* = .15. Table 12 depicts correlations and confidence intervals.

Table 12.

Beery VMI Correlations

			PSI-Diff	CD-Diff	SS-Diff
Beery VMI	**Pearson correlation**		**-.15**	**-.16**	**-.12**
	Sig. (1-tailed)		.17	.15	.23
	95% CI	Upper	-.41	-.43	-.39
		Lower	.12	.10	.16

Note: CI = Confidence interval
PSI-Diff = PSI-D – PSI-P; CD-Diff = CD-D – CD-P; SS-Diff = SS-D – SS-P
Sig. = significance

Results of the Pearson correlations indicated there was not a significant association between scores on the Beery VMI and scores on the PSI-D *r*(39) = .17, *p* = .281 and was a significant positive association between scores on the Beery VMI and scores on the PSI-P *r*(39) = .36, *p* = .020. There was no significant difference between correlation coefficients *z* = 0.89; *p* = 0.186 (one tailed). Results of the Pearson correlations indicated there was not a significant association between scores on the Beery VMI and scores on SS- D *r*(39) = .18, *p* = .255, and was a marginal positive association between scores on the Beery VMI and scores on SS-P *r*(39) = .30, *p* = .053. There was no significant difference between correlation coefficients *z* = .57, *p* = .284 (one tailed). Results of the Pearson correlations indicated there was not a significant association between scores on the Beery VMI and scores on CD-D *r*(39) = .13, *p* = .397, and was a significant positive association between scores on the Beery VMI and scores on CD-P *r*(39) = .31, *p* = .042. There was no significant difference between correlation coefficients *z* = .84, *p* = .200 (one tailed). Table 13 depicts aforementioned results.

Table 13.

Beery VMI Correlations and Differences

	Beery VMI			
	Correlation(r)	Sig. (1-tailed)	Difference (z)	Sig. (1-tailed)
PSI-D	.17	.281		
PSI-P	**.36**	**.020**		
Difference			.89	.186
SS-D	.18	.255		
SS-P	.30	.053		
Difference			.57	.284
CD-D	.13	.397		
CD-P	**.31**	**.042**		
Difference			.84	.200

Sig. = significance; **bolded = Significant correlation**

It was hypothesized scores on the Beery VMI would negatively correlate with the PSI-Diff, CD-Diff, and SS-Diff scores. It was also hypothesized scores on the Beery VMI would positively correlate with scores on paper-pencil versions of the PSI, CD, and SS, to a greater degree than with scores on digital versions of the PSI, CD, and SS. Results did not support the aforementioned hypotheses, thus an exploratory analysis was conducted. Data from participants with scores at or below one standard deviation below the mean (i.e., a standard score of 85 or lower) on the Beery VMI were examined. Data from seven participants were reviewed. Two participants (28.6%) performed over 15 points (i.e., one standard deviation) higher on the PSI-D than the PSI-P and over 3 points (i.e., one standard deviation). No additional meaningful differences were observed. In sum, results of the exploratory analysis did not indicate participants in this sample with below average scores on a measure of visual-motor coordination (i.e., the Beery VMI) performed better on digital versions of processing speed measures than paper-pencil versions (See Table 14).

Table 14.

Beery VMI Exploratory Analysis

	VMI	PSI-D	PSI-P	PSI-Diff	CD-D	CD-P	CD-Diff	SS-D	SS-P	SS-Diff
1	84	119	95	24	15	11	4	12	7	5
2	81	129	135	-6	14	14	0	16	18	-2
3	83	98	95	3	11	11	0	8	7	1
4	84	95	100	-5	8	11	-3	10	9	1
5	83	98	98	0	10	10	0	9	9	0
6	81	83	86	-3	4	5	-1	10	10	0
7	84	141	114	27	15	11	4	19	14	5

Chapter IV

Discussion

This study explored the relationship between paper-pencil and digital formats of processing speed measures from the WISC-V: the PSI, CD, and SS. The impact of psychomotor coordination on differences in performance between paper-pencil and digital versions of subtests was also examined. Participants included 41 students from a private school in New Jersey between the ages of 13.0 and 16.11. Participants were administered a demographic questionnaire and the Beery VMI full form followed by CD-D, CD-P, SS-D, and SS-P subtests in a counterbalanced sequence.

Three main findings were apparent. First, results indicated correlations between paper-pencil and digital scores on CD, SS, and the PSI were lower than thresholds for equivalence (.70 standard set for equivalence by Pearson; .79 standard set for the PSI based on the PSI correlation between scores on WISC-V-P and WAIS-IV-P). Second, the correlation between SS-D and CD-D (*r* = .67) was higher and the correlation between SS-D and SS-P (*r* = .61) was lower than expected. Third, examinees' performances on measures of visual-motor coordination did not share a significant relationship with differences between performances on digital format and paper-pencil format.

Implications of Research Findings

Digital and paper-pencil formats of processing speed subtest scores yielded correlations similar to correlations from previous studies. As expected, CD, SS, and the PSI correlations were lower than hypothesized thresholds for equivalence (.70 for SS and CD; .70 and .79 for the PSI). The correlation between scores on SS-D and SS-P in this sample (*r* = .61) is slightly lower than the correlation from the original equivalence study (*r* = .67; Raiford & Zhang, et al., 2016). The correlation between scores on CD-D and CD-P in this sample (*r* = .67) is comparable to the correlation from the original equivalence study (*r* = .63; Raiford & Zhang, et al., 2016).

The correlation between scores on the PSI-D and the PSI-P was not made available in the original equivalence study (Raiford & Zhang, et al., 2016). Analogous correlations for comparison include the correlation between scores on the WISC-V PSI-P and the WISC-IV PSI-P (*r* = .70) and the correlation between scores on the WISC-V PSI-P and the WAIS-IV PSI-P (*r* = .79). As hypothesized, the correlation between scores on the PSI-D and the PSI-P in this sample (*r* = .68) is lower than the aforementioned PSI correlations. One might expect scores intended for use interchangeably within the same battery (i.e., the WISC-V PSI-P and the WISC-V PSI-D) to correlate higher than scores sharing similar constructs across different batteries (e.g., the WISC-V PSI and the WAIS-IV PSI). Similarly, correlations between scores on the the WISC-V CD-P and the WAIS-IV CD-P (*r* = .69) and scores on the WISC-V CD-P and the WISC-IV CD-P (*r* = .69) correlate higher than scores on the WISC-V CD-D and the WISC-V CD-P (*r* = .67) in the current study. The correlation between scores on the WISC-V SS-P and the WAIS-IV SS-P (*r* = .61) is as high as the correlation between scores on the WISC-V SS-D and the WISC-V SS-P in the current study (*r* = .61). Table 13 depicts information on correlations.

Scores on a measure do not correlate higher with scores on a different measure than they do with themselves over a short time period. Test-retest reliability coefficients from scores on the WISC-V-P ($r_{\text{CD-P}}$ = .81, $r_{\text{SS-P}}$ = .80, $r_{\text{PSI-P}}$ = .83) are consistent with coefficients from scores on the WISC-V-D ($r_{\text{CD-D}}$ = .80, $r_{\text{SS-D}}$ = .75). Given test-retest reliability coefficients of the WISC-V measures are between .75 and .83, it is surprising to observe correlations ranging from .61 to .68 between measures intended for equivalence (i.e., paper-pencil and digital formats of processing speed measures). Table 15 depicts information on test-retest reliability coefficients.

Table 15.

Correlation and Test-Retest Reliability Coefficients

			PSI	CD	SS
Norming Sample(s)	Correlations	WISC-V-P/WISC-V-D	-	.63	.67
		WISC-V-P/WAIS-IV-P	.79	.69	.61
		WISC-V-P/WISC-IV-P	.70	.69	.54
		WISC-V-P/WPPSI-IV	.34	-	-
	Reliability Coefficients	WPPSI-IV	**.84**	**-**	**-**
		WISC-V-P	**.83**	**.81**	**.80**
		WISC-V-D	-	**.80**	**.75**
		WISC-IV-P	**.86**	**.87**	**.78**
		WAIS-IV-P	**.87**	**.86**	**.81**
Study Sample		WISC-V-P/WISC-V-D	.68	.67	.61

Note. Compiled from Wechsler 2014 and Raiford & Zhang, et al., 2016
Correlations between tests
Test-retest reliability coefficients

Correlations between scores on paper-pencil and digital formats of CD (r = .67), SS (r = .61), and the PSI (r = .68) are lower than .70, the standard set for equivalence by Pearson (Wechsler, 2014). However, upper confidence intervals, critical to consider as there is error in all measurement, exceed .70 for CD (CI [.45, .81]), SS (CI [.39, .77]) and the PSI (CI [.49, .82]). This observation may be attributable to the small sample size. Smaller sample sizes yield larger confidence intervals and make it more difficult to detect small differences.

The Multitrait Multimethod Matrix (MTMM) is used to examine evidence for convergent and discriminant validity of measures (Campbell & Fiske, 1959). Within MTMM, variance in each measure is partly due to the construct and partly due to the method. For example, variance in SS-D is partly due to the construct being measured (searching for symbols within an array) and partly due to the method (digital format). Tests designed to measure the same construct (within construct) should correlate higher than tests sharing the same method (within method). Correlations between CD and SS in paper-pencil and digital formats are depicted in Table 16.

Table 16.

MTMM Matrix

	CD-P	SS-D
CD-D	**.67**	.67
SS-P	.48	**.61**

Correlations within construct
Correlations within measure

Results of the MTMM were mixed. The correlation between scores on scores on CD-P and scores on CD-D (*r* = .67) is higher than the correlation between scores on CD-P and scores on SS-P (*r* = .48), as expected. The correlation between scores on SS-D and scores on CD-D (*r* = .67) is higher than the correlation between scores on SS-D and scores on SS-P (*r* = .61). The correlation between scores on CD-D and scores on CD-P (*r* = .67) is equivalent to the correlation between scores on CD-D and scores on SS-D (*r* = .67). The correlation between scores on SS-P and scores on CD-P (*r* = .48) is lower than the correlation between scores on SS-P and scores on SS-D (*r* = .61), as expected. This may indicate the method (digital administration) is contributing more variance in SS scores than desired; SS scores may not reflect the construct intended to be measured. Correlations between subtests measuring different constructs in the same format are not available for comparison as data was not included in the original equivalence study (Raiford & Zhang, et al., 2016). Based on MTMM results, CD-P appears interpretable. Scores on SS may not reflect the construct intended to be measured and digital administration of subtests may be contributing more variance than desired.

Difference scores were created to isolate discrepancies between methods by subtracting scores on paper-pencil formats of processing speed measures from scores on digital formats of processing speed measures (PSI-Diff = PSI-D - PSI-P; CD-Diff = CD-D – CD-P; SS-Diff = SS-D – SS-P). It was hypothesized that scores on the Beery VMI, a measure used to assess visual-motor coordination, would negatively correlate with difference scores because of the psychomotor component present in paper-pencil administration. It was also hypothesized scores on the Beery VMI would positively correlate with scores on paper-pencil versions of the PSI, CD, and SS, to a greater degree than with scores on digital versions of the PSI, CD, and SS. Scores on the Beery VMI and scores on paper-pencil versions of the PSI and CD shared significant positive correlations. The correlation between scores on the Beery VMI and scores on the PSI-P was not significantly different from the correlation between scores on the Beery VMI and scores on the PSI-D. Additionally, the correlation between scores on the Beery VMI and scores on CD-P was not significantly different from the correlation between scores on the Beery VMI and scores on CD-D. The null hypotheses were retained. Correlations between scores on the Beery VMI and scores on the PSI-P and scores on CD-P were significant; additional correlations and difference scores were nonsignificant.

Raiford and Zhang, et al. (2016) conducted a special group study with children with significant motor impairment. Children with motor impairment were administered digital versions of processing speed subtests (CD and SS) and verbal comprehension subtests (Vocabulary and Similarities). In previous studies on paper-pencil administration, the mean performances of children with significant motor impairment were within the average range on verbal comprehension subtests and within the below average range on processing speed subtests (Wechsler, 2003). Raiford and Zhang, et al. (2016) hypothesized mean performance on CD-D would be within the average range in the motor impaired group because the psychomotor component of CD was significantly reduced in digital administration compared to paper-pencil administration. Results were similar to previous research (Wechsler, 2003). Mean scores on CD-D, SS-D, and the PSI-D were significantly lower than mean scores of the matched control group;

mean scores on Vocabulary, Similarities, and Verbal Comprehension are not significantly different from mean scores of the matched control group. This suggests children with motor impairment do not benefit from digital administration of processing speed measures, despite the reduced psychomotor component (Raiford & Zhang, et al., 2016).

Despite these findings, Raiford and Zhang, et al. (2016) suggested the difference in graphomotor skill requirement in digital subtests compared to paper-pencil subtests, specifically for Coding, was reduced and meaningful. Researchers recommended removing graphomotor demands from the interpretation of performance on digital subtests. Assuming this recommendation is applicable, data should show a boost in scores in digital administration compared to paper-pencil administration of processing speed subtests with reduced performance on psychomotor tasks. In the current study, it is possible the Beery VMI is not a sensitive enough measure to detect slight differences in psychomotor processing.

Practical Implications

Processing speed correlations between digital and paper-pencil formats were lower than thresholds for equivalence; statistical differences could not be proven. Observed correlation coefficients between scores on paper-pencil and digital versions of CD (*r* = .67), SS (*r* = .61), and the PSI (*r* = .68) are lower than .70 (the standard set for equivalence by Pearson; Wechsler, 2014). This is similar to observed correlations from the original equivalence study (Raiford & Zhang, et al., 2016). However, confidence intervals, critical to consider because there is error in all measurement, exceed .70 for CD (CI [.45, .81]), SS (CI [.39, .77]) and the PSI (CI [.49, .82]). Furthermore, the correlation between scores on the WISC-V PSI-D and the WISC-V PSI-P (*r* = .68) was lower than the correlation between scores on the WISC-V PSI-P and the WAIS-IV PSI-

P (r = .79). It is important to understand, practically speaking, the WISC-V PSI-P scores correlate higher with the PSI scores from a different battery (e.g., the WAIS-IV PSI-P scores) than they do with scores intended for use interchangeably within the same battery (i.e., the WISC-V PSI-D scores).

The correlation between scores on SS-D and on CD-D (r = .67) is higher than expected given correlations between scores on SS-P and on SS-D (r = .61) and between scores on CD-D and on CD-P (r = .67). The conclusion SS-D and CD-D are measuring two different constructs may be unfounded. The correlation within method (i.e., digital administration) is as high as correlation within the construct of CD, is higher than the correlation within the construct of SS, and is near the standard for equivalence by Pearson. SS-D is more related to CD-D than to its counterpart score intended for use interchangeably (i.e., SS-P). Practitioners should exercise caution during interpretation of subtests as they are intended for interpretation in the context of a larger construct (e.g., processing speed and furthermore, intelligence). Specifically, these results show support for CD-P and SS-P; results do not show strong support for evidence of convergent and discriminant validity of CD-D and SS-D. Based on aforementioned equivalence results, it is important for practitioners to disclose the format of the WISC being administered (i.e., paper-pencil or digital) within reports.

Limitations

The generalizability of the current study is limited because participant data was collected from one private school in New Jersey. Participants were predominantly European American students between the ages of 13.0 and 16.11. Thus, results from this study may not be representative of children from other age ranges, ethnicities, and geographic locations. The sample size used was collected from four separate age bands: 13.0-13.11, 14.0-14.11, 15.0-

15.11, and 16.0-16.11. It is possible a tighter age band (e.g., participants between the ages of 14.0 and 14.11) may have yielded different results. While this would have further limited the generalizability of findings, it would have been more consistent with data collection in previous WISC studies (Raiford & Zhang, et al., 2016; Wechsler, 2014).

Since this was a convenience sample, fewer students participated than originally anticipated. A small sample size, similar to the original equivalence study (i.e., between 20 and 49 participants per age group; Raiford & Zhang, et al., 2016), limited the power of the study to detect differences. Finally, it is possible the Beery VMI is not a sensitive enough measure to detect slight differences in visuomotor coordination. The highest score attainable for our age range (13.0-16.11) is 107, less than one standard deviation above the mean. Limited sensitivity of the Beery VMI may have made finding relationships with difference scores challenging.

Future Research

It would be interesting in the future to conduct a study with a larger sample size. The standard for equivalence currently accepted by Pearson researchers is .70; this correlation seems low as score correlated at .70 share only 49% variance. A study with a larger sample size would have adequate power to detect smaller differences between correlations. Additionally, it is important for future research to explore the relationship between SS and CD in digital format; results imply the method (digital administration) may be contributing higher than expected to variance in scores.

Future research should be conducted to assess differences in paper-pencil and digital administration of processing speed measures across age bands. It is possible that younger children have more significant variability in psychomotor capabilities. Early adaptation of digital measures of processing speed highlighted difficulties in the 6.0-7.11 age band (Raiford & Zhang, et al., 2016). Pilot studies allowed examinees to respond to items on digital administration with a stylus. Due to observed difficulties with grip in the 6.0-7.11 age band, use of a stylus was excluded from standardized administration for all age bands (Raiford & Zhang, et al., 2016).

Each participant in this study had access to an iPad that she or he was able to use in school. iPads and interaction with digital platforms are becoming more commonplace in educational settings. While Pearson poses previous experience interacting with an iPad is not necessary, it would be interesting to look at cohort differences across age bands. This could be accomplished once the WAIS-V is released, assuming processing speed subtests will be released in digital format at that time. It may be possible to compare data collected from participants with little to no interaction with digital platforms to participants with frequent digital interaction.

Conclusions

Findings from this study suggest digital and paper-pencil formats of processing speed subtest scores correlated similarly to results from previous studies (e.g., Raiford & Zhang, et al., 2016). The correlation between digital and paper-pencil formats of WISC-V processing speed index scores is lower than should be expected based on correlations with processing speed index scores from different batteries (i.e., the WISC-IV PSI-P and the WAIS-IV PSI-P). Correlations between scores on paper-pencil and digital formats of CD, SS, and the PSI did not reach .70, the standard set for equivalence by Pearson. However, the upper ranges of confidence intervals exceeded the .70 threshold. Evidence of convergent and divergent validity for processing speed subtests was also mixed. The correlation between scores on subtests sharing the same method (i.e., CD-D and SS-D) was higher than expected. This raises the possibility digital administration is contributing to variance in scores more than intended. Finally, examinees' performances on measures of visual-motor coordination did not share a significant relationship with differences between performances on digital format and paper-pencil format. In sum, there is not enough evidence for practitioners or researchers to interpret SS-D and CD-D subtests scores, or to use scores from digital and paper-pencil methods interchangeably.

References

Beery, K. E., Buktenica, N. A., & Beery, N. A. (2010). *The Beery-Buktenica Developmental Test of Visual-Motor Integration: Administration, scoring, and teaching manual* (6th ed.). Minneapolis, MN: Pearson.

Bradley, J. (1958). Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association, 53*(282), 525-528. doi:10.2307/2281872

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, *54*(1), 1

Cohen, J. (1988). Statistical power analysis for the behaviors science.(2nd). *New Jersey: Laurence Erlbaum Associates, Publishers, Hillsdale*.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Daniel, M. H. (2012a). Equivalence of Q-interactive administered cognitive tasks: WAIS-IV. *Q-interactive Technical Report 1*. Bloomington, MN: Pearson.

Daniel, M. H. (2012b). Equivalence of Q-interactive administered cognitive tasks: WISC-IV. *Q-interactive Technical Report 2.* Bloomington, MN: Pearson.

Daniel, M. H. (2012c). Equivalence of Q-interactive and paper administrations of cognitive tasks: CVLT-II and selected D-KEFS subtests. *Q-interactive Technical Report 3.* Bloomington, MN: Pearson.

Daniel, M. H. (2013). Equivalence of Q-interactive and paper scoring of academic tasks: Selected WIAT-III subtests. *Q-interactive Technical Report 5.* Bloomington, MN: Pearson.

Daniel, M. H., Wahlstrom, D., & Zhang, O. (2014). Equivalence of Q-interactive and paper administration of cognitive tasks: WISC-V. *Q-interactive Technical Report 7*. Bloomington, MN: Pearson.

Daniel, M. H., Wahlstrom, D., & Zhou, X. (2014). Equivalence of Q-interactive and paper administrations of language tasks: Selected CELF-5 tests. *Q-interactive Technical Report 7*. Bloomington, MN: Pearson.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227-246).

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281-306.

Horn, J. L., & Blankson, N. (2005). Foundations for Better Understanding of Cognitive Abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 41-68). New York, NY, US: Guilford Press.

Jacobson, L. A., Ryan, M., Martin, R. B., Ewen, J., Mostofsky, S. H., Denckla, M. B., & Mahone, E. M. (2011). Working Memory Influences Processing Speed and Reading Fluency in ADHD. *Child Neuropsychology*, *17*(3), 209–224. http://doi.org/10.1080/09297049.2010.532204

Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, *47*(7), 635-650

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices. Third Edition.* New York: Springer

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical Manual. *Woodcock-Johnson IV.* Rolling Meadows, IL: Riverside.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation, second edition.* Los Angelas, CA: Sage.

Raiford, S. E., Holdnack, J., Drozdick, L., & Zhang, O. (2014). Q-interactive special group studies: The WISC-V and children with intellectual giftedness and intellectual disability. *Q-interactive Technical Report 11.* Bloomington, MN: Pearson.

Raiford, S. E., Drozdick, L., & Zhang, O. (2015). Q-interactive special group studies: The

WISC- V and children with autism spectrum disorder and accompanying language impairment or attention-deficit/hyperactivity disorder. *Q-interactive Technical Report 11.* Bloomington, MN: Pearson.

Raiford, S. E., Drozdick, L. W., & Zhang, O. (2016). Q-interactive special group studies: The WISC-V and children with specific learning disorders in reading or mathematics. *Q-interactive Technical Report 13.* Bloomington, MN: Pearson.

Raiford, S. E., Zhang, O., Drozdick, L. W., Getz, K., Wahlstrom, D., Gabel, A., Holdnack, J. A., & Daniel, M. (2016). WISC-V coding and symbol search in digital format: Reliability, validity, special group studies, and interpretation. *Q-interactive Technical Report 12.* Bloomington, MN: Pearson.

Sattler, J, M. (2008). *Assessment of children cognitive foundations, fifth edition.* La Mesa, CA: Jerome M. Sattler, Publisher, Inc.

Sattler, J. M., Dumont, R., & Coalson, D. L. (2016) *Assessment of children WISC-V and WPPSI-IV.* La Mesa, CA: Jerome M. Sattler, Publisher, Inc.

Shanahan, M. A., Pennington, B. F., Yerys, B. E., Scott, A., Boada, R., Willcutt, E. G., Olson, R. K., & DeFries, J. C. (2006). Processing Speed Deficits in Attention Deficit/Hyperactivity Disorder and Reading Disability. *Journal Of Abnormal Child Psychology*, *34*(5), 584-601.

Wechsler, D (2003). *Wechsler intelligence scale for children (4th ed.).* Bloomington, MN: Pearson.

Wechsler, D. (2014). *Wechsler intelligence scale for children (5th ed.).* Bloomington, MN: Pearson.

Zhu, J. J., & Chen, H-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29*(6), 570-580.