Journal of School Psychology 85 (2021) 113–124

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

An independent examination of the equivalence of the standard and digital administration formats of the Wechsler Intelligence Scale for Children-5th Edition

Kacey Gilbert [a,*], John H. Kranzler [b], Nicholas Benson [c]

[a] Eastern Washington University, United States of America
[b] University of Florida, United States of America
[c] Baylor University, United States of America

ARTICLE INFO

Action Editor: Eric Buhs

Keywords:
Intelligence
Cognitive abilities
Q-interactive
WISC-V
Equivalence

ABSTRACT

A new administration format for the Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V; Wechsler, 2014) was introduced in 2016 on Q-interactive, Pearson's digital platform for test administration and scoring. The current study examined the measurement unit equivalence of the WISC-V standard and digital administration formats using counter-balanced administration of the 10 primary subtests to measure intellectual ability. The results indicated that correlations (*r*) between standard scores on subtests and composites administered in each format were generally moderate, with mean *r*s of 0.64 for subtests and 0.71 for composites after correction for attenuation, with the lowest *r*s for processing speed. Split-plot ANOVAs were conducted to examine within-subjects main effects for administration format and order and their interaction. The results of these analyses revealed significant main effects for format for the Full Scale IQ and Processing Speed composite scores, with small to medium effect sizes ($ds > 0.40$). These format effects largely stemmed from the non-equivalence of the Coding subtest, which is used to derive both composites. For Coding, the main effect for format was statistically significant, with a large effect size ($d = 0.69$). Statistically significant administration order by format interaction effects were also observed for a number of composites and subtests, with medium to large effect sizes ($\eta_p s^2 > 0.20$). In each case, higher mean scores were observed when the WISC-V was administered first in digital format. Implications of these results for research and practice are discussed.

The Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949, 1974, 1991, 2003, 2014a) has long been one of the most used psychological instruments by school psychologists (Benson et al., 2019; Goh et al., 1981; Hutton et al., 1992; Reschly et al., 1987; Stinnett et al., 1994; Wilson & Reschly, 1996). In a recent survey of test use and assessment practices, Benson et al. (2019) found that the most recent edition of the WISC, the Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V; Wechsler, 2014a), was the second most administered instrument overall and that 80% of all school psychologists reported using it within the past year. Moreover, they found that school psychologists administered the WISC-V more frequently than the next five most used tests of intelligence *combined*. The WISC-V was administered an average of 3.5 times per month ($SD = 4.8$), whereas the next most frequently administered intelligence test, the Differential Ability Scales-Second Edition (DAS-2; Elliot, 2007), was given less than once per month, on average.

* Corresponding author at: Psychology Department, College of Social Sciences, Eastern Washington University, 135 Martin Hall, Cheney, WA 99004, United States of America.
E-mail address: kgilbert5@ewu.edu (K. Gilbert).

https://doi.org/10.1016/j.jsp.2021.01.002
Received 19 May 2020; Received in revised form 18 December 2020; Accepted 25 January 2021
Available online 27 February 2021

According to these survey results, the WISC-V is clearly the "gold standard" for the assessment of intelligence in school psychology.

Each edition of the WISC has differed significantly from its predecessor, and the WISC-V is no exception. In the development of the WISC-V, two core subtests were removed (i.e., Word Reasoning and Picture Completion) from its predecessor (Wechsler Intelligence Scale for Children-Fourth Edition [WISC-IV]; Wechsler, 2003) and three new subtests were added (i.e., Figure Weights, Picture Span, and Visual Puzzles) for the measurement of intellectual ability. Moreover, new items were created, or existing items modified, on all of the WISC-IV subtests that were retained. Other revisions included updated norms, new index scores, major changes to score terminology, and modifications to administration procedures (Wechsler, 2014b). Further, the test developers stated that by "incorporating new research on intelligence, cognitive development, neurodevelopment, cognitive neuroscience, and processes important to learning, the WISC-V is distinct from its predecessor" (Wechsler, 2014b, p.1).

In addition to major changes to its content and structure, in 2016 a new administration format for the WISC-V was introduced on Q-interactive, Pearson Inc.'s digital platform for test administration and scoring. The Q-interactive system requires the use of two Apple iPads that are connected wirelessly via Bluetooth. The examiner's iPad serves the same functions as the WISC-V manual and response form. This iPad, however, not only presents items and records responses (including audio), but also registers response times and score responses, among other benefits (e.g., alerting the examiner when a discontinue rule has been met, automatic conversion of raw scores to scaled scores). On subtests with visual stimuli, the examiner's iPad is used to send images for each item to the examinee's tablet. The examinee's iPad essentially serves as a digital stimulus book, and on some subtests, the examinee's iPad records answers selected by screen touches and automatically sends them back to the examiner's tablet for programmed scoring (after review). All primary subtests on the standard version of the WISC-V have been adapted for administration in digital format, except Block Design. On this subtest, the visual stimuli are presented to the examinee via their iPad, but the traditional red-and-white blocks are still used for responding. In addition to these noteworthy enhancements to WISC-V administration, the vast majority of examiners using Q-interactive tend to report that children and youth tend to be more engaged and motivated during intelligence testing when administration is done digitally (Daniel, 2013).

1. Equivalence of WISC-V administration formats

The digital version of the WISC-V was not developed, normed, and validated as a new standalone instrument. Rather, it was adapted to the original WISC-V. As Daniel and Wahlstrom (2019) stated, "when digital tests are adaptations of paper tests[1]…publishers are obligated to show whether the norms and other psychometric information based on the original paper versions are applicable to the digital versions" (p. 1). In other words, research evidence must demonstrate that scores on the standard and digital versions of the WISC-V are *equivalent*. Daniel et al. (2014) discussed a number of possible ways in which the digital administration of the WISC-V may threaten equivalence, including those related to examiner and examinee interaction with the iPad (e.g., differences in presentation of stimuli and response requirements), accuracy of the Q-interactive response capture and scoring, and global effects of the digital assessment environment (e.g., young children perceiving the iPad as a toy).

There are two main types of test equivalence (e.g., Van de Vijver & Poortinga, 2005). The first is *construct equivalence*. A construct is a characteristic of individuals, such as intelligence, which is assumed to be reflected in test performance (Cronbach & Meehl, 1955). Construct equivalence occurs when different versions of a test measure the same underlying construct. Without construct equivalence, there is no basis for comparisons across different versions of a test. Demonstrating this form of equivalence requires both internal and external evidence. One of the most important and widely used statistical techniques to substantiate the tests' internal structure is factor analysis, of which there are two types: exploratory and confirmatory (e.g., Thompson, 2004). The two most widely used methods to establish external evidence of test equivalence are correlations with external criteria and group differentiation. The main goal of this kind of research is to compare the pattern of relations (convergent and discriminant) between the constructs measured by the different forms of the test and other measures of similar and dissimilar constructs or other salient outcome criteria (see Messick, 1995). Composite and subtest scores must show similar patterns of correlations with other variables and differences between divergent groups (e.g., clinical and non-clinical) across administration formats to substantiate construct equivalence.

The second type of equivalence is *measurement unit equivalence*.[2] Measurement unit equivalence is observed when the different versions of a test have the same distribution of scores. When scores for different administration formats are on the same metric (e.g., age-based standard scores), they should yield the same means and standard deviations. Examination of measurement unit equivalence commonly involves the use of an equivalent groups or retest design (Daniel, 2014). In the equivalent groups design, a representative sample of the population is selected and participants are randomly assigned to one of two groups, each of which takes the test once in a different format. In contrast, in the retest design, each participant takes the test twice. They are randomly assigned to one of two administration order groups and given the tests in a counter-balanced order. According to Daniel and Wahlstrom (2019), measurement unit equivalence can be seen as evidence of construct equivalence when the digital version (a) closely resembles the standard version in terms of the mode of stimulus input and response processes, and (b) produces the same distribution of scores.

What evidence supports the equivalence of the standard and digital administration formats of the WISC-V? A number of studies have examined the measurement unit equivalence of psychological tests that have been adapted for the Q-interactive platform,

[1] Daniel and Wahlstrom (2019) use "paper" here broadly to describe all tests administered in non-digital format.

[2] Measurement unit equivalence is also referred to in the literature as *raw-score equivalence*. Although Daniel and Wahlstrom (2019) used the term raw-score equivalence in their study, they did not examine the equivalence of raw scores, but age-based scaled scores. Because we also used scaled scores in our study, we used the term measurement unit equivalence to avoid any confusion about the type of scores used.

including the WISC-IV and the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008). According to Daniel (2012, 2014), scores on all 15 subtests of the WAIS-IV and on 13 of 15 subtests of the WISC-IV were equivalent across the standard and digital administration formats. The two subtests that were found to be non-equivalent on the WISC-IV were Matrix Reasoning and Picture Concepts. Nevertheless, given the extensiveness of the revisions in the most recent edition, the generalizability of these results to the WISC-V is questionable. To date, only two technical reports have been disseminated by Pearson, Inc. on the equivalence of the standard and digital administration formats of the WISC-V (Daniel et al., 2014; Raiford et al., 2016) and only one was subsequently published in a peer-reviewed journal (Daniel & Wahlstrom, 2019).

Daniel and Wahlstrom (2019) examined the measurement unit equivalence of 18 of the 21 subtests of the WISC-V (cf. Daniel et al., 2014). The primary processing speed subtests (i.e., Coding, Symbol Search, and Cancellation) were not included due to the fact that preliminary research failed to establish their measurement unit equivalence. In an equivalent groups design, Daniel and Wahlstrom recruited 350 children and youth from the general population between the ages of 6–16 years, excluding those with perceptual and motor disabilities and other clinical conditions. They randomly assigned participants to one of two administration format groups ($ns = 175$). Each group was administered the 18 WISC-V subtests in either standard or digital format. After the completion of testing, Daniel and Wahlstrom stated that they matched all participants in both groups by age, gender, race/ethnicity, and parent education, resulting in 175 matched-pairs that were approximately representative of the general child population. Although demographic data for the two groups were not presented by Daniel and Wahlstrom, they were included in their original technical report (see Daniel et al., 2014). In this original report, the two groups did not have the same number of participants for all demographic variables, thereby indicating that not all 175 pairs were matched on the four demographic variables (see Table 1, p. 8). Thus, the two groups were not "equivalent," although they were quite comparable. In addition, the high degree of similarity in the number of participants for each demographic variable across groups suggests that stratified sampling was used to recruit participants and not random sampling from the general population.

Daniel and Wahlstrom (2019) examined the measurement unit equivalence of the standard and digital administration formats of the WISC-V by conducting a simultaneous multiple regression analysis for each of the 18 subtests. In these analyses, the subtest's scaled score was the dependent variable (DV) and the independent variables (IVs) were (a) the four demographic variables, (b) a dummy-coded variable for administration format (Standard = 0, Digital = 1), and (c) a "selection of WISC-V subtests/indexes" (p. 4) found to have small format effects in an equivalence study of the WISC-IV (Daniel, 2012). The unstandardized regression weight for the format variable was divided by 3 (the subtest scaled score standard deviation) to calculate the effect size of the administration format. Daniel and Wahlstrom defined an acceptable effect size for equivalence for clinical use as 0.20 or less (cf. Cohen, 1988).[3] The results of their analyses revealed effect sizes below this cut-off ranging from 0.02 to 0.20, with a mean of 0.11. In addition, they calculated correlations to examine whether the size of the difference in administration format was related to age, gender, race/ethnicity, and parent education, as well as "ability level" (i.e., the predicted standard-administration score for that subtest). Results of these analyses revealed negligible correlations between format effects and the demographic variables and ability level. Daniel and Wahlstrom (2019) concluded that "results of this study indicate few differences between the paper and Q-interactive versions of the WISC-V" (p. 5). Although these results are suggestive of equivalence at the scale level, it is important to note that the authors did not examine equivalence at the item level. Researchers commonly use an item response theory (IRT) model to examine differential item functioning (e.g., von Davier, 2013). As this was not done, the extent to which individual items are equivalent across administration formats is unclear.

In a technical report, Raiford et al. (2016) provided the results of several studies examining the equivalence of the Coding and Symbol Search subtests of the WISC-V across administration format. Cancellation was not included because preliminary research found that it could not be adapted for digital format. Preliminary research also found that measurement unit equivalence for Coding and Symbol Search could not be established. Consequently, Raiford et al. used the inferential scaling method (Zhu & Chen, 2010) to carry over the scale from the standard administration format to the digital format. Thus, although scaled scores from both formats are on the same metric, raw score to scale score conversions differ between formats. For the inferential scaling study, Raiford et al. used stratified sampling to recruit 329 children and youth between the ages of 6–16 years, approximately representative of the general population in terms of age, gender, race/ethnicity, parent education level, and geographic region. Participants were administered the digital Coding and Symbol Search subtests in the recommended subtest administration order along with the other eight primary subtests. After testing was completed, these two processing speed subtests were then administered in the paper format.

After using inferential scaling to place age-based scaled scores for the standard and digital administration formats on the same metric, Raiford et al. (2016) evaluated the construct equivalence of the two processing speed tests across these formats. In the first study, they examined the test-retest reliability of scaled scores on the digital Coding and Symbol Search subtests using two groups, one aged 6–7 years ($n = 33$) and the other aged 8–16 years ($n = 35$). Although these samples were rather small, they were diverse in terms of gender, race/ethnicity, parent education level, and geographic region. The mean interval between first and second testing was 25.3 days, with an interval range of 14–72 days. The resultant stability coefficients were 0.74 for Coding and 0.77 for Symbol Search for the 6- to 7-years-olds, and 0.85 for Coding and 0.78 for Symbol Search for the 8- to 16-year-olds. These results indicated that the digital Coding and Symbol Search subtests had at least adequate stability and were comparable to those for the standard administration format (Wechsler, 2014b).

[3] According to Daniel (2014), the researchers at Pearson Inc. "selected an effect size of .20 as the smallest effect that would be a threat to the use of Q-interactive results interchangeably with scores from paper-format administration" (p. 2). An effect size of 0.20 is equivalent to 0.6 scaled-score points on subtests ($M = 10$, $SD = 3$) and 3.0 points on composites ($M = 100$, $SD = 15$) on the WISC-V.

Raiford et al. (2016) also examined patterns of relations with other variables and the internal structure of the WISC-V with the digital processing speed subtests. For these studies, they used stratified sampling to recruit 651 children and youth between the ages of 6–16 years, approximately representative of the general population in terms of age, gender, parent education level, and geographic region. For race/ethnicity, however, students from Hispanic backgrounds were slightly over-represented in the sampling. The standard and digital formats of the Coding and Symbol Search subtests were administered in a counter-balanced order. According to Raiford et al., half of the participants within each stratum were randomly assigned to take the digital format first in the regular subtest administration order, followed by administration of the other subtests in the standard format. The other half of the participants were administered the subtests by format in the opposite order. For each group, participants were given the Coding and Symbol Search subtests in one format in the recommended subtest administration order along with the other eight primary subtests in standard format, followed by administration of the processing speed subtests in the other format.

Raiford et al. (2016) conducted correlations for all subtest and composite scaled scores on the WISC-V for the overall sample and for groups of children aged 6-to-7 years and children and youth aged 8-to-16 years. Raiford et al. did not report demographic data by group. Although not mentioned in the technical report, it appears that the digital test scores were pooled across administration order for these analyses. Subtests that comprise an index should correlate more highly with each other than they do with other subtests comprising other indices, and because the Processing Speed subtests tend to have lower loadings on psychometric *g*, their correlation with the FSIQ should be lower than other primary subtests. Raiford et al. found that not only were correlations between the Coding and Symbol Search subtests in digital format higher with each other than they were with the other primary subtests, but both subtests correlated substantially with the Processing Speed index and only moderately with the Full Scale IQ (FSIQ). Thus, the results indicated that the pattern of correlations for the overall sample and for both age groups were as predicted and consistent with construct theory.

Using these same data, Raiford et al. (2016) conducted a number of confirmatory factor analyses (CFAs) to compare the internal structure of the WISC-V with the digital processing speed subtests and without. For each CFA, they examined the goodness-of-fit of the data to a model based on the theoretical structure of the WISC-V (i.e., five first-order factors corresponding to the primary index scores, and one second-order general factor corresponding to the FSIQ). For the overall sample and within each age group, they conducted two CFAs. In the first, they examined the fit of the 10 standard primary subtests administered in standard administration format. In the second, they substituted the two digital processing speed subtests for the standard ones and examined model fit again. The results revealed that goodness-of-fit was very good for all analyses within each group, regardless of the format. Assessing model fit in CFA, however, includes more than examining goodness-of-fit indices. It also requires examining the factor solution. Although Raiford et al. only presented the factor model for the overall sample with the digital subtests, the parameter estimates for the processing speed subtests were of the appropriate sign and size. The digital processing speed subtests also had salient but moderate loadings on psychometric *g* for the overall sample, as predicted by construct theory.[4]

In addition to these correlational analyses, Raiford et al. (2016) reported descriptive statistics for the raw and scaled scores for the Coding and Symbol Search subtests for both administration formats, although only for the overall sample. They did not report descriptive statistics by age group or by administration order within groups, nor did they report the results of within-subjects statistical analyses to examine the main effects for administration format and order and their interaction. They did report the effect size of the mean differences across administration formats, however. Although the effect size for Symbol Search of $d = 0.13$ was below their cut-off effect size for clinical use of 0.20, the effect size for Coding of $d = 0.23$ exceeded it. They also reported correlations between administration formats. Correlations for raw scores were 0.87 for Coding and 0.84 for Symbol Search, and scaled scores were 0.60 for Coding and 0.67 for Symbol Search. In summary, despite the fact that one effect size exceeded their cut-off for clinical use, Raiford et al. (2016) concluded that "these results provide strong support that the subtests are measuring similar constructs whether in digital or paper format" (p. 16).[5]

2. Critique of the Pearson Inc. technical reports

One of the main shortcomings of the studies by Daniel and Wahlstrom (2019, cf. Daniel et al., 2014) and Raiford et al. (2016) is that neither examined the equivalence of the standard and digital administration formats of all primary WISC-V subtests used to measure intellectual ability in the same study. Daniel and Wahlstrom examined the equivalence of all subtests, except those measuring processing speed; and Raiford et al. only examined the equivalence of two processing speed subtests (Coding and Symbol Search). More important, neither study examined the equivalence of *any* of the primary indices, or the FSIQ, which has been found to be the most

[4] Although the goodness-of-fit indices for the overall sample with the digital processing speed subtests were very good, the correlation between the second-order Full Scale factor and the first-order Fluid Reasoning factor was 0.99, indicating that these two factors are functionally indistinguishable. These results are consistent with independent research suggesting that the WISC-V is over-factored (Canivez, Dombrowski, & Watkins, 2018; Canivez et al., 2016, 2017; Dombrowski et al., 2018), although some disagree (cf. Reynolds & Keith, 2017).

[5] The technical report by Raiford et al. (2016) also provided the results of a number of clinical group differentiation studies with the Coding and Symbol Search subtests in digital format. Group differentiation studies are important at the external stage of construct validation to establish the meaning of test scores by appraising the degree to which differences between certain groups (e.g., groups with intellectual disability or giftedness and a non-clinical sample) are consistent with construct theory (Messick, 1995). Nevertheless, while important for establishing the meaning of test scores, clinical group differentiation studies shed no light on their equivalence across the traditional and digital administration formats. Because our study focused solely on the measurement unit equivalence of the WISC-V administration formats, readers are referred to the technical report for information on the results of their group differentiation studies.

reliable and valid score on the WISC-V (Kranzler & Floyd, 2020). This is a significant omission, especially given that many researchers and test developers currently recommend that the primary level of clinical interpretation of the WISC-V be done at the index score level (e.g., Flanagan & Alfonso, 2017; Kaufman, Raiford, & Coalson, 2016; Wechsler, 2014a).

In addition to this limitation, Daniel and Wahlstrom (2019) did not employ a within-subjects design. Although they claimed to have used an "equivalent groups design," not only were their groups in fact not equivalent, but they used multiple regression and controlled for the demographic variables and "ability level" statistically. In addition, if they used stratified sampling to recruit subjects, which seems likely, then the demographic variables in their regression analyses are fixed effects and not random effects. When data are not drawn independently from the population, the assumption that errors (residuals) are independent is violated. Although violation of this assumption will not affect regression coefficients, it will affect errors. Consequently, the results of the tests of statistical significance of the IVs in Daniel and Wahlstrom's regression analyses – which included a dummy-coded variable for administration format – are questionable.

In contrast, Raiford et al. (2016) administered the Coding and Symbol Search subtests in both digital and standard formats in a counter-balanced design. Counter-balancing administration enhances internal validity by providing maximum control of extraneous participant variables. Participants in all conditions have the same age, gender, socioeconomic status, cognitive ability, and so on, because they are the very same people. Moreover, within-subjects designs tend to have more power than between-subjects designs (e.g., Bausell & Li, 2002). The primary disadvantage of this design is that it can result in "carryover" effects (e.g., Cotton, 1993). A carryover effect is an effect that "carries over" from one experimental condition to another. The performance of participants in later conditions may either be improved or worsened due to the order in which the conditions were presented (e.g., practice or fatigue effects). Whenever subjects perform in more than one condition (as they do in within-subject designs) there is a possibility of carryover effects. Although counter-balancing is based on the assumption that order effects are the same regardless of the specific sequence of conditions, it is important to note that it does not remove possible confounds due to carryover effects. The presence of potential confounding carryover effects, however, can be detected by conducting within-subjects analyses, such as split-plot ANOVA. This is especially important to do before pooling data across conditions. Given that Raiford et al. did not report the results of such within-subjects analyses, they presumably did not conduct them, despite pooling their data for the digital subtests across administration order. Thus, due to the possible confounding of carryover effects, results of the construct equivalence studies reported in their technical report are also questionable.

3. Purpose of the current study

Scientific evidence must demonstrate that scores on the standard and digital administration formats are equivalent prior to the digital administration of the WISC-V on Q-interactive. At the current time, the internal validity of research that has been conducted on the equivalence of the standard and digital administration formats is open to question (Daniel et al., 2014; Daniel & Wahlstrom, 2019; Raiford et al., 2016). In addition, no independent research has been conducted on the equivalence of WISC-V administration formats. The purpose of this study, therefore, was to examine the measurement unit equivalence of the WISC-V standard and digital administration formats using counter-balanced administration of the 10 primary subtests used to measure intellectual ability. Relations between subtests and composites were examined across administration format. In addition, within-subjects analyses were conducted to examine the equivalence of scaled scores for the primary subtests and index scores, as well as the FSIQ.

Table 1
Demographic characteristics of participants by administration format.

	Administration order	
	Standard-digital ($n = 31$)	Digital-standard ($n = 34$)
Sample	*n* (%)	*n* (%)
Gender		
Boys	11 (36%)	13 (38%)
Girls	20 (64%)	21 (62%)
Race/Ethnicity		
White	26 (84%)	24 (79%)
Black	2 (7%)	4 (12%)
Hispanic	3 (9%)	4 (12%)
Other	0 (0%)	2 (6%)
Lunch Status		
Free/Reduced	2 (6%)	5 (15%)
Full	29 (94%)	29 (85%)
	M (*SD*)	M (*SD*)
Age in Years	9.6 (2.3)	10.1 (2.8)
Test-Retest Interval in Weeks	5.5 (1.4)	5.7 (1.6)

4. Method

4.1. Participants

Participants were 65 children and youth recruited from public (62%), charter (29%), and private (8%) schools in Florida ($n = 53$), Tennessee ($n = 3$) and Texas ($n = 9$). One additional participant in Florida was homeschooled (<1%). Those with perceptual and motor disabilities and other clinical conditions were excluded from participation. Approximately half of the participants who were recruited from public schools attended a K-12 developmental research school in North Central Florida. The mission of this school is to serve as a vehicle for research, demonstration, and evaluation regarding teaching and learning. Its primary role is to design, develop, evaluate, and disseminate exemplary programs of education. Because the generalizability of results is an important concern for researchers, students at the school are selected so that the student body is diverse and comparable to that of the state's school-age population in terms of gender, race/ethnicity, socioeconomic status (SES), and academic achievement.

4.2. Instrument

The WISC-V is a standardized, norm-referenced instrument that was developed to assess the intellectual abilities of children and youth between the ages of 6–16 years (Wechsler, 2014a). Intellectual ability on the WISC-V is measured by 10 primary subtests. These subtests are used to derive the five primary index scores (Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed). These index scores are "factor-based and recommended for a comprehensive description and evaluation of intellectual ability" (Wechsler, 2014b, p. 7). The FSIQ, a hierarchical factor-based score, is an overall score that is derived from 7 of the 10 primary subtests. Six secondary subtests can be administered to provide a more comprehensive assessment of intellectual ability. In addition, if one of the seven subtests used to derive the FSIQ is missing or invalid, a specific secondary subtest can be used as a substitute. No substitution is allowed for derivation of the primary index scores. Different combinations of the primary and secondary subtests can be used to derive five ancillary index scales (Quantitative Reasoning, Auditory Working Memory, Nonverbal, General Ability, and Cognitive Proficiency). Finally, five complementary subtests can be administered to derive complementary index scales (Naming Speed, Symbol Translation, and Storage and Retrieval) for special clinical uses, such as analysis of patterns of strengths and weaknesses for the identification of specific learning disabilities (e.g., Flanagan & Alfonso, 2017; Kranzler et al., 2020). The complementary subtests are used only to derive the complementary index scores and cannot be substituted for a primary subtest to derive the intellectual ability index scores.

For the current study, only the 10 primary subtests on the WISC-V were administered, because these are the subtests used to derive the primary index scores and the FSIQ. In addition, the ancillary and complimentary index scores are not factor-based scores and empirical substantiation of their validity is currently lacking. Information on the development of the WISC-V and the reliability and validity evidence supporting its use can be found in the *Technical and Interpretive Manual* (Wechsler, 2014b) and in reviews (e.g., Canivez & Watkins, 2016).

4.3. Devices

The Apple iPads used in this study were the same version as those used in the equivalence studies by Daniel et al. (2014) and by Raiford et al. (2016). These were 6th Generation iPads with 9.7″ screens. Apple pencils (1st Generation) were used to record the responses and take notes. To ensure connectivity between the examiner and examinee's iPads, all tablets were configured according to the guidelines provided by Pearson Inc. (2020).

4.4. Procedures

Examiners were doctoral students in school psychology training programs (six in Florida, one in Texas). All examiners were Caucasian women in their 20s to 40s who were in their second year of training or beyond. All had completed coursework in intellectual assessment and had experience in administering the WISC-V in standard format. They received additional training on the digital administration of the WISC-V on the Q-interactive platform, which included viewing training videos provided by Pearson, Inc. Prior to the study, all examiners demonstrated competence by completing several practice digital test administrations. The WISC-V was administered in both standard and digital formats under standardized conditions (Wechsler, 2014a). After random assignment to group, the WISC-V was administered to each participant twice in a counter-balanced design.

The principal of each school in which data were collected facilitated recruitment by sending consent forms to parents via email. Students returning signed consent forms were randomly assigned to one of two groups. One group was administered the WISC-V in the standard format first, followed by administration in the digital format ($n = 31$). The other group was administered the WISC-V in the opposite format order ($n = 34$). The abrupt closure of schools and social distancing mandates due to the coronavirus pandemic led to the abrupt cessation of data collection, resulting in unequal group sample sizes. Table 1 presents demographic information for participants by group. As can be seen here, the two groups were roughly equivalent in terms of age, gender, race/ethnicity, and free or reduced lunch status. For the overall sample, the mean age was 9.8 years ($SD = 2.6$). Also shown in Table 1 are descriptive statistics for the interval between test administrations. For the overall sample, the mean interval between test administrations was 5.6 weeks ($SD = 1.5$).

4.5. Data analysis

Descriptive statistics were calculated for all variables in both groups. Zero-order correlations were conducted to examine the relations between subtest and composite age-based scaled scores across administration formats. To examine the main hypotheses of the study, a series of split-plot ANOVAs were conducted to examine the main effects for administration format and order, as well as the interaction between administration format and order, on subtest and composite scaled scores. An alpha level of 0.05 was used for all statistical tests. The measure of effect size in the split-plot analyses was the partial eta-squared (η_p^2), which is the amount of variance in the DV that is explained after partialling out the effects of other IVs and interactions. According to Cohen et al. (2003), $\eta_p^2 = 0.01$, 0.09, and 0.25 can be classified as "small," "medium," and "large" effects, respectively.

5. Results

Table 2 displays descriptive statistics for the scaled scores on the primary subtests ($M = 10$, $SD = 3$) and composites ($M = 100$, $SD = 15$) of the WISC-V by administration format. As can be seen here, mean standard scores for all indexes and the FSIQ for both formats are within the average to above average range. For the FSIQ, the mean for the standard format was $M = 110.5$ ($SD = 10.9$) and $M = 113.3$ ($SD = 12.1$) for the digital administration format. In addition, there was some restriction of range for the composites, with the exception of Processing Speed for the digital format ($SD = 16.7$). Descriptive statistics for the primary subtests were generally similar to those for composites. The above average mean performance was not surprising, given that the majority of participants were recruited from local private schools and a developmental laboratory research public school. Students enrolled in these schools are not randomly selected from the population; rather, parents must apply for their child's admission. Hence, this may reflect some self-selection on cognitive ability. Table 2 also displays the results of within-sample *t*-tests for each composite and subtest score. As can be seen here, statistically significant mean differences were observed for the FSIQ, the Proceesing Speed index, and the Coding subtest ($ps < 0.05$). According to Cohen (1988) $|ds| = 0.20$, 0.50, and 0.80 can be classified as "small," "medium," and "large," respectively. Using this rule-of-thumb, the effect sizes for significant results shown in Table 2 are small to medium, but exceed Pearson Inc.'s 0.20 cut-off for clinical use (cf. Daniel & Wahlstrom, 2019).

Table 3 shows Pearson product-moment correlations (*r*) between scaled scores for the primary subtests and composites across administration format. Given that the range of scores was generally less than that commonly observed in the population as a whole, the *rs* after correction for attenuation (Wiberg & Sundström, 2009) are also shown. As can be seen in Table 3, the corrected *r* for the FSIQ was 0.89. For the primary indexes, corrected *rs* ranged from 0.62 to 0.87, with $M = 0.71$. For subtests, corrected *rs* ranged from 0.50 to 0.83, with $M = 0.64$. The highest *rs* were for the Verbal Comprehension index and the subtests from which it is derived (Similarities and Vocabulary). In contrast, the lowest *rs* were for the Processing Speed index and the subtests used to derive it (Coding and Symbol Search), as well as for the Picture Span subtest. Many of the correlations for the subtests were moderate and only two exceeded 0.70.

Table 4 displays the results of split-plot ANOVAs conducted to examine main effects for administration format and order and the interaction of format by order. As be seen here, for the composites, statistically significant main effects for format were observed for both the FSIQ ($F = 13.03$, $p < .05$, $\eta_p^2 = 0.17$) and the Processing Speed index ($F = 14.19$, $p < .05$, $\eta_p^2 = 0.18$), with medium to large effect sizes. The main effect of administration order was also statistically significant for the Visual Spatial index ($F = 4.82$, $p < .05$, $\eta_p^2 = 0.07$), with a small to medium effect size. In addition, the administration format by order interaction effect for the FSIQ was

Table 2
Descriptive statistics by administration format.

	Administration Format ($N = 65$)						
	Standard			Digital			
Index Scaled Score	M	SD	Range	M	SD	Range	*d*
Full Scale IQ	110.5	10.9	82–132	113.3	12.1	91–139	0.41*
Verbal Comprehension	112.6	14.2	78–150	113.1	13.5	86–146	0.06
Fluid Reasoning	108.9	13.5	76–137	110.3	12.8	82–140	−0.09
Working Memory	109.6	13.8	79–146	108.7	13.1	82–138	0.12
Visual Spatial	109.7	10.8	84–129	108.9	13.4	72–138	−0.09
Processing Speed	104.0	12.1	72–141	109.8	16.7	75–141	0.45*
Subtest Scaled Score							
Similarities	12.1	3.3	3–19	12.6	2.9	7–19	0.22
Vocabulary	12.5	2.6	5–18	12.3	2.7	5–18	−0.12
Block Design	11.6	2.2	6–16	11.4	2.9	4–17	−0.10
Visual Puzzles	12.0	2.4	8–19	11.8	2.4	6–17	−0.09
Matrix Reasoning	11.2	2.7	5–17	11.3	2.9	5–17	0.04
Figure Weights	12.0	2.9	2–19	11.9	3.1	0–19	−0.02
Digit Span	11.0	2.8	6–18	11.2	2.8	5–16	0.06
Picture Span	12.3	3.0	5–19	11.9	2.9	4–19	−0.13
Coding	10.0	2.6	3–16	12.1	3.6	4–19	0.69*
Symbol Search	11.3	2.7	4–18	11.0	2.9	5–17	−0.05

* $ps < 0.05$.

Table 3
Pearson product-moment correlations (r_{xy}) between administration formats.

Index Scaled Scores	r_{xy}	Corrected r_{xy}[a]
Full Scale IQ	0.83	0.89
Verbal Comprehension	0.85	0.87
Fluid Reasoning	0.64	0.69
Working Memory	0.68	0.72
Visual Spatial	0.64	0.72
Processing Speed	0.60	0.62
Subtest Scaled Scores		
Similarities	0.80	0.79
Vocabulary	0.80	0.83
Block Design	0.63	0.69
Visual Puzzles	0.55	0.64
Matrix Reasoning	0.56	0.59
Figure Weights	0.56	0.56
Digit Span	0.67	0.70
Picture Span	0.49	0.50
Coding	0.57	0.56
Symbol Search	0.49	0.52

[a] Wiberg and Sundström (2009). All $ps < 0.05$.

Table 4
Split-plot ANOVA results.

Index Scaled Score	Order	Format	Interaction
Full Scale	$F = 1.89$ $\eta_p^2 = 0.03$	$F = 13.03$* $\eta_p^2 = 0.17$	$F = 8.41$* $\eta_p^2 = 0.12$
Verbal Comprehension	$F = 0.00$ $\eta_p^2 = 0.00$	$F = 0.25$ $\eta_p^2 = 0.00$	$F = 0.00$ $\eta_p^2 = 0.00$
Fluid Reasoning	$F = 2.07$ $\eta_p^2 = 0.03$	$F = 1.33$ $\eta_p^2 = 0.02$	$F = 8.32$* $\eta_p^2 = 0.12$
Working Memory	$F = 0.45$ $\eta_p^2 = 0.01$	$F = 0.394$ $\eta_p^2 = 0.01$	$F = 3.15$ $\eta_p^2 = 0.08$
Visual Spatial	$F = 4.82$* $\eta_p^2 = 0.07$	$F = 0.40$ $\eta_p^2 = 0.01$	$F = 67.10$* $\eta_p^2 = 0.52$
Processing Speed	$F = 0.23$	$F = 14.19$*	$F = 2.61$
	$\eta_p^2 = 0.00$	$\eta_p^2 = 0.18$	$\eta_p^2 = 0.04$
Subtest Scaled Score			
Similarities	$F = 0.16$ $\eta_p^2 = 0.00$	$F = 3.30$ $\eta_p^2 = 0.05$	$F = 0.16$ $\eta_p^2 = 0.00$
Vocabulary	$F = 1.17$ $\eta_p^2 = 0.02$	$F = 0.69$ $\eta_p^2 = 0.01$	$F = 0.92$ $\eta_p^2 = 0.01$
Block Design	$F = 0.52$ $\eta_p^2 = 0.01$	$F = 0.52$ $\eta_p^2 = 0.01$	$F = 23.20$* $\eta_p^2 = 0.27$
Visual Puzzles	$F = 2.71$ $\eta_p^2 = 0.04$	$F = 0.43$ $\eta_p^2 = 0.01$	$F = 22.29$* $\eta_p^2 = 0.26$
Matrix Reasoning	$F = 2.40$ $\eta_p^2 = 0.04$	$F = 0.21$ $\eta_p^2 = 0.00$	$F = 5.50$* $\eta_p^2 = 0.08$
Figure Weights	$F = 2.13$ $\eta_p^2 = 0.03$	$F = 0.01$ $\eta_p^2 = 0.00$	$F = 5.62$* $\eta_p^2 = 0.08$
Digit Span	$F = 0.60$ $\eta_p^2 = 0.01$	$F = 0.27$ $\eta_p^2 = 0.00$	$F = 0.27$ $\eta_p^2 = 0.00$
Picture Span	$F = 0.92$ $\eta_p^2 = 0.00$	$F = 1.01$ $\eta_p^2 = 0.02$	$F = 3.93$ $\eta_p^2 = 0.06$
Coding	$F = 0.67$ $\eta_p^2 = 0.01$	$F = 30.39$* $\eta_p^2 = 0.33$	$F = 1.70$ $\eta_p^2 = 0.03$
Symbol Search	$F = 0.36$ $\eta_p^2 = 0.01$	$F = 0.92$ $\eta_p^2 = 0.00$	$F = 17.12$* $\eta_p^2 = 0.21$

* $ps < 0.05$.

statistically significant ($F = 8.41$, $p < .05$, $\eta_p^2 = 0.12$), with a medium to large effect size. This interaction is ordinal and resulted from a higher mean score when the WISC-V was administered first in the digital format. Finally, statistically significant format by order interaction effects were observed for the Fluid Reasoning ($F = 8.32$, $p < .05$, $\eta_p^2 = 0.12$) and Visual Spatial ($F = 67.10$, $p < .05$, $\eta_p^2 = 0.52$) indexes, with medium and large effects, respectively. For both of these indexes, the significant interaction effects were

disordinal and resulted from higher mean scores for the digital administration format on the first testing occasion.

For the subtests, results of split-plot analyses revealed that the only statistically significant main effect for administration format was for Coding ($F = 30.39$, $p < .05$, $\eta_p^2 = 0.33$), with a large effect size. Mean scores for the digital administration format were higher, on average, than those for the standard format on both testing occasions. In addition to this main effect, statistically significant interaction effects were observed for the Block Design ($F = 23.20$, $p < .05$, $\eta_p^2 = 0.27$), Visual Puzzles ($F = 22.29$, $p < .05$, $\eta_p^2 = 0.27$), Matrix Reasoning ($F = 5.50$, $p < .05$, $\eta_p^2 = 0.08$), and Figure Weights ($F = 5.62$, $p < .05$, $\eta_p^2 = 0.08$) subtests, with medium to large effects for each. As was observed for composites, each of the statistically significant administration format by order interaction effects for subtests resulted from higher mean scores for the digital format when it was administered first.

To further examine these statistically significant interaction effects, we calculated difference scores for the group that was administered the digital format first ($n = 31$) by subtracting each participant's score on the standard administration format (given second) from their score on the iPad (given first) for all composites and subtests. We then correlated these difference scores with age, gender, and FSIQ. The results indicated that difference scores were not statistically significantly related to age or gender ($ps > 0.05$), but they were correlated +0.63 ($p < .001$) with FISQ. In addition, 7 out of 31 participants obtained a higher score on the standard version when it was administered after the iPad, but only one with an FSIQ >115 ($n = 15$). These results indicate that the difference between formats when the iPad was administered first was related to general cognitive ability, with larger declines in test scores on the second adminstration in standard format for those with higher general cognitive ability.

6. Discussion

In 2016, Pearson Inc. introduced a new administration format for the WISC-V on Q-interactive, its digital platform for test administration and scoring. Pearson Inc., however, has disseminated only two technical reports on the equivalence of the standard and digital versions of the WISC-V (viz., Daniel et al., 2014; Raiford et al., 2016) and only one was subsequently published in a peer-reviewed journal (i.e., Daniel & Wahlstrom, 2019). Therefore, the purpose of the current study was to examine the measurement unit equivalence of the WISC-V standard and digital administration formats with counter-balanced administration of the 10 primary subtests used to measure intellectual ability.

We first examined the relationships between the composites and subtests across administration format. The mean *r* for composites was 0.75 (Range = 0.62 to 0.87) and 0.64 (Range = 0.49 to 0.80) for subtests, after correction for attenuation. The highest *r*s were for the Verbal Comprehension index and the subtests used to derive it (Similarities and Vocabulary) and the lowest were for the Processing Speed index and the subtests used to derive it (Coding and Symbol Search), in addition to Picture Span. The *r*s of 0.56 for Coding and 0.52 for Symbol Search – which are much smaller than those reported in the *Technical and Interpretive Manual* of the WISC-V (Wechsler, 2014b) – indicate that only about 25% of the variance in one format is accounted for by the other. In contrast, the corrected test-retest reliability coefficients for the standard administration format for the Proceesing Speed index was 0.83, 0.81 for Coding, and 0.80 for Symbol Search. The pattern of correlations in our study undoubtedly reflects the amount of adaption that was required for each subtest. For example, the administration of Similarities and Vocabulary are virtually identical across administration formats and required little adaption. During the digital adaption of the Processing Speed subtests, however, marked changes were made to the manner of stimuli presentation and examinee response requirements (e.g., on-screen touch responses, scrolling stimuli, elimination of writing).

If the standard and digital administration formats of the WISC-V are truly equivalent, then correlations between subtests and composites across administration format in our study should be comparable to test-retest reliability coefficients for the standard administration format. As can be seen in the *Technical and Interpretive Manual* of the WISC-V, however, the corrected stability coefficients for the primary subtests and composites in the standard administration format are uniformly higher, and in some cases much higher than what we found (see Wechsler, 2014b, p.63). For the standard administration format, stability coefficients for the overall sample (ages 6–16) ranged from 0.75 to 0.94, with a $M = 0.85$ for composites, and from 0.78 to 0.90 for subtests, with a $M = 0.82$. Although the mean number of days between test administrations in our study ($M = 39$, Range = 20–67) was somewhat larger, on average, than in the study that examined the stability of the standard administration of the WISC-V ($M = 26$, Range = 9–82), the completely overlapping ranges in the number of days between test administrations and the small mean difference between test-retest intervals suggests that differences in elapsed time between testing occasions is an unlikely explanation of the lower *r*s in our study. Thus, the fact that the *r*s between subtests and composites in our study are substantially lower than the test-retest reliability coefficients for those same variables can be taken as evidence of the non-equivalence of the standard and digital administration formats.

In addition to these findings, results of split-plot ANOVAs revealed statistically significant main effects for administration format for the FSIQ and the Processing Speed index, with small to medium effect sizes. A statistically significant main effect for format was also observed for the Coding subtest, with a medium to large effect size, according to Cohen's criteria (1988). This effect size is larger than the cut-off of 0.20 set by Pearson Inc. for sufficient equivalence for use in the field (see Daniel & Wahlstrom, 2019). Given that the Coding subtest is one of two subtests used to derive the Processing Speed index and one of seven used to derive the FSIQ, the statistically significant main effects for administration format for the Processing Speed and FSIQ composites stem in large part from the measurement unit non-equivalence of Coding.

On February 4, 2020, in an email to Q-interactive customers, Pearson acknowledged the non-equivalence of the administration formats for Coding and stated "that the level of inflation rose at the end of 2019 in relation to a baseline period in 2018" and that "further analyses showed the common thread connecting elevated scores was administration using newer model iPad® devices." The results of our study, however, were obtained using the same, older iPads as those used by Daniel et al. (2014) and Raiford et al. (2016) in their equivalence studies. Based on our results, it seems unlikely that the observed non-equivalence of the administration formats of the WISC-V is due to model of iPad per se.

Instead, a more plausible explanation is that non-equivalence results from the inferential scaling process used to scale the raw score to scaled score conversions for the Coding and Symbol Search subtests (see Zhu & Chen, 2010). Despite its advantages, this method is particularly susceptible to errors when small sample sizes are used. As Zhu and Chen stated, when it comes to the development of inferential norms, "the larger the sample size, the better the curve estimates" (p. 578). Using data from the WISC-IV, Zhu and Chen determined that "$N = 50$ per age group could provide norms with decent qualities, but should be considered the lowest acceptable sample size" (p. 579). The sample used by Raiford et al. (2016), however, consisted of 329 children and youth for 11 one-year age groups between the ages of 6–16 years. Their scaling study, therefore, was based on only $n \approx 30$ per age group, on average, which is far below Zhu and Chen's minimum recommended sample size. Thus, it seems likely that the non-equivalence of the WISC-V administration formats results from flawed inferential scaling.

Results of a recent meta-analysis of retest effects on tests of cognitive ability (Scharfen et al., 2018) revealed that retesting tends to result in an average increase of approximately one-third of a standard deviation in test scores on the second administration and that these increases are not related to general cognitive ability. Pearson Inc. also reported retesting increases of a similar magnitude on the WISC-V for the standard administration format (Wechsler, 2014b). The results of our study, however, revealed only one statistically significant administration order effect – for the Visual Spatial index – and in the opposite direction (i.e., decrease in mean test scores on the second administration). In addition, we found a number of statistically significant administration format by order interaction effects, with medium to large effect sizes. These interactions were mainly observed for subtests that involved reasoning with non-verbal visual materials. Specifically, these were Block Design, Visual Puzzles, Matrix Reasoning, and Figure Weights – the subtests comprising the Visual Spacing and Fluid Reasoning indexes – but the interaction for Symbol Search was also significant. In each of these instances, the significant result stemmed from examinees performing better, on average, on the digital format when it was administered first. Results of analyses of difference scores across format for the group administered the digital format first indicated that the size of these differences was positively correlated with general cognitive ability. Our findings, therefore, suggest that the effects of examinee engagement and motivation when the WISC-V is administered in digital format may be more nuanced than previously believed (e.g., see Daniel, 2013).

7. Future research directions and limitations

Our study is the first to examine the measurement unit equivalence of the standard and digital administration formats of the WISC-V using all 10 primary tests to measure intellectual ability. Despite Daniel and Wahlstrom's (2019) assertion that measurement unit equivalence can be taken as evidence of construct equivalence when the digital and standard administration formats closely resemble each other and have the same distribution of scores, not only do the digital and standard versions of the WISC-V not closely resemble each other closely for all subtests (i.e., processing speed subtests), but our results indicate that they do not have the same score distributions. Thus, the validity evidence gathered for the standard administration format of the WISC-V does not necessarily apply to the digital format. Instead, further independent research is needed to substantiate the equivalence of the different administration formats and the validity of the digital format.

What additional research is needed? At the current time, the internal structure of the WISC-V has not been examined using all 10 of the primary digital subtests. Although Raiford et al. (2016) used CFA to examine the fit of the data to a model based on the theoretical structure of the WISC-V, they substituted two digital processing speed subtests for the standard ones and examined model fit along with the other eight primary subtests administered in standard format. The stability of the digital administration format also has not been investigated. Moreover, no study using all 10 primary subtests has examined patterns of relations of the digital administration format with external criteria or group differentiation (clinical versus non-clinical), nor has research been conducted using an IRT model to examine the equivalence of differential item functioning across administration formats. Last, but not least, the question of test bias (e.g., gender, socioeconomic status, and race/ethnicity) has not been studied on the digital administration format of the WISC-V. Given that the WISC-V is the most administered intelligence test in school psychology (Benson et al., 2019), further research on the digital administration format is needed.

Furthermore, in a recent review, Wahlstrom et al. (2016) concluded that in comparison to the standard administration format, higher mean scores tend to be obtained on the WISC-V when using Q-interactive on subtests that require the examinee to provide a touch response on the iPad. According to them, no differences across administration formats are typically observed on subtests for which the iPad is either used to display visual stimuli with no touch response or for which the tablet is not used at all. According to the results of our study, however, a potentially important factor related to the digital administration of the WISC-V – administration order – appears to have gone undetected in prior equivalence research. We found only one statistically significant main effect for administration format, with a large effect size, which is consistent with Wahlstrom et al.'s assertion for Coding, but we also found a number of statistically significant administration order by format interaction effects. With the exception of Coding, the results of our study suggest that differences between administration formats are not related solely to the use of Q-interactive on the iPad, but to one or more within-subjects variables that interact with administration format and order (e.g., motivation and/or attention). One possible explanation of these interaction effects may be related to the distinction between typical and maximum performance (e.g., Sackett et al., 1988). Typical performance is how someone performs on a regular basis, whereas maximum performance is how one performs when exerting as much effort as possible. Maximum performance may be related to conative factors and degree of attentional control that are not completely under conscious control (see Schneider & McGrew, 2018). It may be that for participants with higher cognitive ability, greater unconscious engagement and focus explain the higher mean performance when the digital format of the WISC-V was administered first. Further research is needed to better understand these interaction effects and, if replicable, their implications for practice.

The main limitation of our study concerns generalizability. Because of the exclusionary criteria used, our results may not be generalizable to clinical populations. In schools, intelligence tests are primarily administered as part of a comprehensive evaluation for the identification of intellectual disability, specific learning disability, and giftedness (Kranzler & Floyd, 2020). Therefore, further research on the equivalence of the standard and digital administration formats of the WISC-V is needed with these and other clinical populations (e.g., attention deficit hyperactivity disorder) to examine the possibility of group-specific effects. Moreover, it is important to note that the WISC-V is a nationally standardized test that is intended to be used with all children and youth between the ages of 6–16 years in the United States, with certain exceptions (e.g., sensory or motor deficits, non-English language speakers). Because of this, any sample for which the WISC-V has been validated is appropriate for examining the equivalency of the digital and standard administration formats. Our sample was not representative of the general population, however. The overall performance of our sample was approximately two-thirds of a standard deviation above the mean, on average. Hence, further research is also needed to examine the generalizability of our results with non-clinical groups that differ from our sample in terms of geographic region, demographic characteristics (e.g., race/ethnicity), and cognitive ability level.

8. Implications for practice

On June 30, 2020, Pearson Inc. announced that the digital Coding and Symbol Search subtests of the WISC-V were being removed from Q-interactive due to issues concerning non-equivalence. Effective with the content update released on July 14, 2020, the subtests on the Processing Speed index can only be administered using the standard paper response booklets. This means that all of the primary subtests used to measure intellectual ability on the WISC-V can be administered on Q-interactive in the usual digital format, except for Symbol Search and Coding. For those needing to use the results of the WISC-V that were administered in an all-digital format (e.g., in triennial re-evaluations), there are three options available to offset the effect of the non-equivalence of the digital Coding subtest, in descending order of viability:

1. *Substitute Symbol Search for Coding to derive the FSIQ.* If administration was done entirely in the digital format, substitute Symbol Search for Coding when deriving the FSIQ. We did not find a statistically significant main effect for administration format on Symbol Search, which means that it can be interpreted when administered in digital format and substituted for Coding when deriving the FSIQ on Q-interactive.
2. *Interpret the General Ability Index (GAI) instead of the FSIQ.* When the WISC-V was administered entirely in digital format, the GAI can be interpreted instead of the FSIQ as a measure of general cognitive ability, or psychometric *g*. Psychometric *g* is a factor, and as such is a distillate of the tests from which it was derived. According to Jensen (1998), a "good" *g* is estimated from a battery of tests that is diverse in terms information content, mode of stimulus input, and mode of response. Because the GAI is derived from five subtests and does not include subtests from the Processing Speed or Working Memory composites, as does the FSIQ, it is based on a somewhat narrower range of cognitive abilities. Thus, it is not as good an estimate of *g* as the FSIQ. Consequently, interpreting the GAI is a less viable option for deriving the FSIQ than substitution.
3. *Prorate to Derive the FSIQ.* The use of proration is allowed on the WISC-V when one of the seven subtests used to derive the FSIQ is invalid. Because of its non-equivalence, proration can be used without Coding to derive the FSIQ when only the seven primary subtests were administered in digital format. Proration involves multiplying the sum of scaled scores for the other six subtests by 7/6 and then using the prorated sum to derive the FSIQ. This proportionate adjustment is based on the assumption that the score on the subtest that was not administered is the same as the mean of the subtests that were. Proration, however, should be used with caution because of the fact that it may introduce unknown measurement errors, which will generate higher confidence intervals for the FSIQ and increase the chance of misclassification (Zhu et al., 2016).

9. Conclusion

According to Benson et al.'s (2019) recent survey of test use and assessment practices, school psychologists are widely using test scoring via software or online services, such as Q-global. Although they found that practitioners tend to use digital technology less frequently to administer psychological tests than to score them, about two-thirds reported doing so on occasion. We welcome the application of technology. In addition to ease of administration and scoring and reduction of administration and scoring errors, digital assessment has other advantages when assessing children and youth, such as increased examinee engagement and motivation. Although not fully realized at present, computer scoring also has the potential to produce refined index scores (i.e., factor scores) that provide a more precise measurement of the constructs they are intended to assess than current hand-scored methods (see Benson et al., 2016). In conclusion, the use of computer technology to administer and score intelligence tests is clearly the future in school psychology. Advances in test administration and use, however, must be substantiated by independent empirical research before being used in the field to make high-stakes decisions about children and youth, including both the determination of eligibility for special education and related services and the design of interventions.

Author note

We thank Delaney Boss, Kathryn Matthews, Melina Yaraghchi, and Bailey Mungiguerra for assistance with data collection.
We have no known conflict of interest to disclose.

References

Bausell, R. B., & Li, Y. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. Cambridge University Press.

Benson, N., Kranzler, J. H., & Floyd, R. G. (2016). Examining the integrity of measurement of cognitive abilities in the prediction of achievement: Comparisons and contrasts across variables from higher-order and bifactor models. *Journal of School Psychology, 58*, 1–19.

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists: Findings from the 2017 national survey of assessment practices in school psychology. *Journal of School Psychology, 72*, 29–48.

Canivez, G. L., Dombrowski, S. C., & Watkins, M. W. (2018). Factor structure of the WISC-V in four standardization age groups: Exploratory and hierarchical factor analyses with the 16 primary and secondary subtests. *Psychology in the Schools, 55*, 741–769.

Canivez, G. L., & Watkins, M. W. (2016). Review of the Wechsler Intelligence Scale for Children-Fifth Edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Raiford, & D. L. Coalson (Eds.), *Intelligent testing with the WISC-V* (pp. 683–702). Wiley.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler Intelligence Scale for Children-Fifth Edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 28*, 975–986.

Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children-Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 29*, 458–472.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge.

Cotton, J. W. (1993). Latin square designs. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science: 137. Statistics: Textbooks and monographs* (pp. 147–196). Marcel Dekker.

Cronbach, L., & Meehl, P. (1955). Construct validity of psychological tests. *Psychological Bulletin, 52*, 281–302.

Daniel, M., & Wahlstrom, D. (2019). Raw-score equivalence of computer-assisted and paper versions of WISC-V. *Psychological Services*. https://doi.org/10.1037/ser0000295.

Daniel, M. H. (2012). *Equivalence of Q-interactive-administered cognitive tasks: WISC–IV*. Pearson.

Daniel, M. H. (2013). *User survey on Q-interactive examinee behavior*. Pearson.

Daniel, M. H. (2014). *Q-interactive: Choosing sample sizes for equivalency studies*. Pearson.

Daniel, M. H., Wahlstrom, D., & Zhang, O. (2014). *Equivalence of Q-interactive and paper administrations of cognitive tasks: WISC-V*. Pearson.

von Davier, A. A. (2013). Observed-score equating: An overview. In *, 78*. *Psychometrika* (pp. 605–623). https://doi.org/10.1007/S11336-013-9319-3.

Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology, 22*, 90–104.

Elliot, C. D. (2007). *Differential Ability Scales (2nd ed.)*. Harcourt Assessment.

Flanagan, D. P., & Alfonso, V. C. (2017). *Essentials of WISC-V assessment*. Wiley.

Goh, D. S., Teslow, C. J., & Fuller, G. B. (1981). The practice of psychological assessment among school psychologists. *Professional Psychology, 12*, 696–706.

Hutton, J. B., Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review, 21*, 271–284.

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

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC-V*. Wiley & Sons Inc.

Kranzler, J. H., & Floyd, R. G. (2020). *Assessing intelligence in children and adolescents: A practical guide for evidence-based assessment* (2nd ed.). Rowman & Littlefield.

Kranzler, J. H., Maki, K. E., Benson, N. F., Floyd, R. G., & Fefer, S. A. (2020). How do school psychologists interpret intelligence tests for the identification of specific learning disabilities? *Contemporary School Psychology*. https://doi.org/10.1007/s40688-020-00274-0.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741.

Pearson, Inc. (2020). *Q-interactive system requirements*. Pearson.

Raiford, S. E., Zhang, O., Whipple Drozdick, L., 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*. Pearson.

Reschly, D. J., Genshaft, L., & Binder, M. S. (1987). *The 1986 NASP survey: Comparison of practitioners, NASP leadership, and university faculty on key issues*. National Association of School Psychologists.

Reynolds, M. R., & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children-Fifth edition: What does it measure? *Intelligence, 62*, 31–47.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482–486. https://doi.org/10.1037/0021-9010.73.3.482.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan, & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 73–163). The Guilford Press.

Stinnett, T. A., Harvey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12*, 331–350.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. *American Psychological Association*. https://doi.org/10.1037/10694-000.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Erlbaum.

Wahlstrom, D., Daniel, M., Weiss, L. G., & Prifitera, A. (2016). Digital assessment with Q-interactive. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), *WISC-V assessment and interpretation: Scientist-practitioner perspectives* (pp. 347–372). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-404697-9.00011-X.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. Pearson.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children-Revised*. Pearson.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children* (3rd ed.). Pearson.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). Pearson.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). Pearson.

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children* (5th ed.). Pearson.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children (5th ed.): Technical and interpretive manual*. Pearson.

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment Research & Evaluation, 14*, 1–9.

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

Zhu, J., Cayton, T. G., & Chen, H. (2016). Substitution, proration, or a retest? The optimal strategy when standard administration of the WPPSI-IV is infeasible. *Psychological Assessment, 28*, 1441–1451. https://doi.org/10.1037/pas0000272.

Zhu, J., & Chen, H. (2010). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29*, 570–580.