Equivalence of Q-interactive™ and Paper Scoring of Academic Tasks: Selected WIAT®–III Subtests

Q-interactive Technical Report 5

Mark H. Daniel, PhD
Senior Scientist for Research Innovation

August 2013

Pearson

Introduction

Q-interactive™, a Pearson digital platform for individually administered tests, is designed to make assessment more convenient and accurate, provide clinicians with easy access to a large number of tests, and support new types of tests that cannot be administered or scored without computer assistance.

With Q-interactive, the examiner and examinee use wireless tablets that are synched with each other, enabling the examiner to read administration instructions, time and capture response information (including audio recording), and view and control the examinee's tablet. The examinee tablet displays visual stimuli and captures touch responses.

In the initial phase of adapting tests to the Q-interactive platform, the goal has been to maintain raw-score equivalence between standard (paper) and digital administration and scoring formats. If equivalence is demonstrated, then the existing norms, reliability, and validity information can be applied to Q-interactive results.

This is the fifth Q-interactive equivalence study. In this study, the equivalence of scores from digitally assisted and standard scorings of two subtests of the *Wechsler Individual Achievement Test*®—third edition (WIAT®–III; Wechsler, 2009), Oral Reading Fluency and Sentence Repetition, were evaluated.

In the first two equivalence studies, all fifteen *Wechsler Adult Intelligence Scales*®–fourth edition (WAIS®–IV; Wechsler, 2008) subtests and thirteen of fifteen *Wechsler Intelligence Scales for Children*®–fourth edition (WISC®–IV; Wechsler, 2003) subtests yielded comparable scores in the Q-interactive and standard (paper) administration formats. On two WISC–IV subtests (Matrix Reasoning and Picture Concepts), scores were slightly higher with Q-interactive administration. The third study evaluated four *Delis-Kaplan Executive Function* Scale™ (D-KEFS™; Delis, Kaplan, & Kramer, 2001) subtests and the *California Verbal Learning Test*®–second edition (CVLT®–II; Delis, Kramer, Kaplan, & Ober, 2000) Free-Recall trials, all of which demonstrated equivalence across digital and paper formats. In the fourth study, three subtests of the NEPSY®–second edition (NEPSY®–II; Korkman, Kirk, & Kemp, 2007) and two subtests of the *Children's Memory Scale*™ (CMS™; Cohen, 1997) were found to be equivalent.

In all the equivalence studies, it is assumed that digitally assisted (Q-interactive) administration may affect test scores for a number of possible reasons, including the following:

- Examinee interaction with the tablet. To minimize effects of examinee–tablet interaction that might threaten equivalence, physical manipulatives (e.g., CMS Dot Locations grid) and printed response booklets (e.g., D-KEFS Trail Making) were used with the Q-interactive administration. Though these physical components may be replaced, eventually, by interactive digital interfaces, the degree of adaptation required could cause a lack of raw-score equivalence. More extensive development efforts would then be required to support normative interpretation and provide evidence of reliability and validity.

- Examiner interaction with the tablet, especially during response capture and scoring. Most of the administration differences in the first version of Q-interactive occurred in the examiner interface. Administering a test on Q-interactive is different from the standard administration because Q-interactive includes tools and procedures designed to simplify and support the

examiner's task. Great care was taken to ensure that these adaptations did not diminish the accuracy with which the examiner presents instructions and stimuli, monitors and times performance, and captures and scores responses.

- Global effects of the digital assessment environment. Global effects go beyond just the examinee's or examiner's interaction with the tablet. For example, a global effect was observed in an early study in which the examiner used a keyboard to capture the examinee's verbal responses. Examinees appeared to slow the pace of their responses so as not to get ahead of the examiner. Because this could lower their scores, the use of a keyboard for response capture was abandoned.

In the Q-interactive studies, if a task was not equivalent across the two formats, the cause of the digital effect was investigated. Understanding the cause is critical to deciding how to deal with the format effect. In principle, if it were determined that Q-interactive makes examiners more accurate in their administration or scoring, then Q-interactive provides an advance in assessment technology, and a lack of equivalence would not necessarily be a problem. A reasonable objective for a new technology is to produce results equivalent to those from examiners who use the standard paper format *correctly.* The digital format should not replicate administration or scoring errors that occur in the standard format. On the other hand, if it appears that a digital effect is due to a reduction in accuracy on the part of either the examinee or the examiner, then the first priority is to modify the Q-interactive system to remove the source of error. Only if that were not possible would the effect be dealt with through norms adjustment.

It is imperative that equivalence studies incorporate a method of checking the accuracy of administration, recording, and scoring in both digital and standard formats. Only then can score discrepancies be attributed to one format or the other, or to particular features of either format. All or most of the Q-interactive equivalence study administrations were video recorded to establish the *correct* score for each item and subtest. These recordings also showed how examiners and examinees interacted with the test materials in each format.

As a whole, the equivalence studies indicate that examinees ages 5 and older (the youngest individuals tested) who do not have a clinical diagnosis or special-education classification respond in a similar way when stimuli are presented on a digital tablet rather than a printed booklet, or when their touch responses are captured by the screen rather than through examiner observation. The one exception (Matrix Reasoning and Picture Concepts) suggests that on subtests involving conceptual reasoning with visual stimuli (or close visual analysis of those stimuli), children may perform better when the stimuli are shown on the tablet; the reason for this difference is not yet known. Also, the cumulative evidence shows that when examiners use the kinds of digital interfaces that have so far been studied in place of a paper record form, administration manual, and stopwatch, they obtain the same results.

Equivalence Study Designs

Several experimental designs have been employed in Q-interactive equivalence studies. In most of them, each examinee takes a subtest only once, in either digital or standard (paper) format. This approach avoids any changes in the way an examinee interacts with the task as a result of having done it before. Ideally, we are trying to detect any effects that the format may have on how the examinee interacts with the task when they encounter it for the first time. Study designs in which there is only a single administration to each examinee provides a realistic testing experience.

The WAIS–IV and WISC–IV studies relied primarily on an *equivalent-groups* design, with either random or nonrandom assignment of examinees to groups. This design compares the performance of two groups, one taking the test in the digital format and the other in the paper format. The equivalent-groups design is described in detail in Q-interactive Technical Reports 1 and 2.

A second design, *retest,* was used in the follow-up study of the WAIS–IV Processing Speed subtests (Technical Report 1) and the study of NEPSY–II and CMS subtests (Technical Report 4). Each examinee takes the subtest twice, once in each format (in counterbalanced order). When a retest design is possible, it is highly efficient because examinees serve as their own controls. This design is appropriate when the response processes are unlikely to change substantially on retest because the examinee does not learn solutions or new strategies for approaching the task or solving the problem.

The third type of design, a single-administration design called *dual-capture*, was used for the CVLT–II and D-KEFS studies (Q-interactive Technical Report 3). This method is appropriate when the digital format affects how the examiner captures and scores responses, but the format is not expected to affect examinee behavior. Each of a relatively small number of examinees takes the test only once, but the administration is video recorded from the examiner's perspective so that it can be viewed by a number of scorers who score it using either paper or digital format. A comparison of average scores with the two formats indicates whether the format affects the response-capture and scoring process. This is the design that was used for the current study of the WIAT–III Oral Reading Fluency and Sentence Repetition subtests.

For all equivalence studies, an effect size of 0.2 or smaller has been used as the standard for equivalence. Effect size is the average amount of difference between scores on Q-interactive and paper administrations, divided by the standard deviation of scores in the population. An effect size of 0.2 is equal to three points on the standard-score metric used for WIAT–III subtests (mean of 100 and standard deviation of 15).

Selection of Participants

The Q-interactive equivalence studies (including this one) have used samples of nonclinical examinees to maintain focus on estimating the presence and size of any effects of the digital format. Because the possible effects of computer-assisted administration on individuals with particular clinical conditions are not known, the inclusion of examinees with various disorders in the sample could obscure the results. Understanding the interaction of administration format with clinical conditions is ultimately of importance for clinical applications of Q-interactive; however, the initial research focuses on the primary question of whether or not the digital format affects scores obtained by nonclinical examinees.

The amount of demographic control required for the sample depends on the type of design. In the equivalent-groups designs, it is important that the samples being compared represent the general population (gender, ethnicity, and socioeconomic status [education level]) and that the two groups are demographically similar to each other. In retest and dual-capture designs, which focus on within-examinee comparisons, examinee characteristics are less significant; however, it is important for the sample to have enough diversity in ability levels and response styles to produce varied responses so that the different features of the digital interface can be evaluated.

Examiners participating in the equivalence studies were trained in the subtests' standard paper administration procedures. Examiners received enough training and practice in the digital administration and scoring procedures to be able to conduct the administration and capture

responses smoothly, without having to devote a great deal of attention to the format. Experience suggests that becoming thoroughly familiar and comfortable with a new format takes at least three practice administrations.

WIAT–III Equivalence Study

Measures

The WIAT–III is a comprehensive, individually administered measure of the academic achievement of preschool and school-age children and adults. The two WIAT–III subtests chosen for this study met two criteria:

- Their Q-interactive interfaces have features that could plausibly affect the examiner's ability to capture and score responses accurately, but which have not previously been studied and are not already known to be free of effects on equivalence.
- During the administration, there is minimal examiner feedback or intervention while the examinee is responding to an item or trial. On Oral Reading Fluency, it is true that the examiner intervenes by saying the word if the student hesitates or struggles for 5 seconds; however, that intervention was considered too little a threat to equivalence to rule out the dual-capture study design.

The two subtests included in this study are:

Oral Reading Fluency. This subtest measures the student's rate and accuracy of reading expository and narrative text aloud, with comprehension. The student reads two grade-level passages and the examiner records the time for each passage as well as errors, which include additions, substitutions (including mispronunciations), omissions, and transpositions. At the end of each passage, the examinee answers a comprehension question that does not affect scoring, but is included to motivate the examinee to read for meaning. The examiner uses the tablet as a timer, and marks errors on an image of the stimulus text in the same way he or she would on a paper record form. (As with the paper administration, the examiner's version of the passage has extra spacing between the lines to accommodate markup of errors.) After administration, the examiner tallies the number of errors and enters that information, using a counter on the screen. There are three scores: Fluency (number of words read correctly per minute), Accuracy (number of words read correctly, minus number of errors), and Rate (total time).

Sentence Repetition. This task set is a portion of the Oral Expression subtest. It measures oral syntactic knowledge and short-term memory. The student hears a sentence and repeats it. The examiner marks errors on an image of the sentence shown on the examiner tablet, and indicates the score for each item (2, 1, or 0) according to the number of errors (additions, omissions, substitutions, or transpositions).

Method

Preliminary Readability Study

Because the reading passages for Oral Reading Fluency were presented on the examinee's tablet, a preliminary study was done to verify that readability was not affected by the tablet presentation. Such an effect was considered unlikely because each reading passage in the printed stimulus booklet was small enough to fit on the tablet screen, with no scrolling required. The screen

dimensions were 5 ¾" by 7 ¾". The width of the text on the printed page was 6", and for all but one of the sixteen passages, shrinking the image by about 5% (to accommodate the screen width) was sufficient. One passage was tall enough (8 ½") to require shrinkage by about 10%.

The study also included two other WIAT–III subtests, Word Reading and Pseudoword Decoding, because they also use printed stimulus pages of text that had to be slightly reduced in size to fit on the tablet screen. For the digital presentations of these two subtests, examinees moved from the first to the second stimulus page by swiping, rather than by flipping the card over as in the paper presentation. The primary score for each of these subtests is the number correct, but the rate (number of items completed in the first 30 seconds) is reported as a supplemental score without norms.

The readability study was conducted in February, 2013 at Pearson's office in San Antonio, TX. The convenience sample included 12 examinees:

- 11 students in grades 4–12 and one adult, age 34
- 7 females and 5 males
- 10 white and 2 other ethnicity
- all with parent (or self) education level of 16+ years

Each examinee took Word Reading (WR), Pseudoword Decoding (PD), and Oral Reading Fluency (ORF), and in that sequence. Easy ORF passages were used for two reasons: first, to ensure the study would be sensitive to a possible format effect on the ability of examinees to read quickly; and second, to enable the analysis to be done using raw scores. Examinees in grades 4–6 took the Grade 3 ORF passages; those in grades 7 and higher took the passages for Grades 7–8. Half of the examinees took WR and the first ORF passage on paper, and PD and the second ORF passage on the tablet; the other half used the reverse sequence of presentation formats. Examinees taking the two sequences were balanced by grade and sex.

The data were analyzed using a retest design, in which WR and PD were treated as alternate forms because both require reading an array of individual words/pseudowords formatted in the same way. The WIAT–III norm tables indicate that the standard deviations of raw scores (number correct) of these subtests are similar in the general population (approximately 12 for Word Reading and 11 for Pseudoword Decoding), and, therefore, the subtests can be treated as alternate forms in the statistical analysis.

The multiple regression analysis of WR/PD used the difference between WR and PD raw scores as the dependent variable, with the predictors being sequence (paper/digital or digital/paper), the average ORF words-per-minute score, the ORF passage level, and demographic characteristics (grade, sex, and ethnicity). The analysis of ORF used the same model, with the difference between the first and second ORF trials being the dependent variable, and with the average WR/PD rate score used as a predictor.

In each analysis, the unstandardized regression weight for sequence, divided by two, is an estimate of the format effect. Table 1 shows the results. None of the estimated format effects were substantial or statistically significant. The effect sizes (in standard deviation units) for the digital presentation were smaller than .20 in absolute value: +.18 for Word Reading/Pseudoword Decoding and –.11 for Oral Reading Fluency. The fact that the effects were in opposite directions adds support to the conclusion that there is no real effect of the digital format on reading rate.

Table 1 Results of readability study

Subtest(s) and Score Type	Format Effect	Effect Size	*p*
Word Reading/ Pseudoword Decoding number correct	2.1	0.18	0.34
Oral Reading Fluency words per minute	–4.2	–0.11	0.45

Note. A positive format effect indicates higher scores with digital presentation of the text stimulus.

Participants

The sample for the main study of Oral Reading Fluency and Sentence Repetition consisted of 10 examinees, 5 examiners, and 9 scorers. (Because Oral Reading Fluency is not administered until Grade 1, the sample for that measure consisted of 8 examinees.) Pearson's Field Research staff recruited examinees and compensated them for their participation. Potential examinees were screened for demographic characteristics and exclusionary factors, such as perceptual or motor disabilities or clinical conditions; except for one examinee with a gifted/talented designation, none of the examinees had special-education classifications or clinical diagnoses. The sampling plan called for approximately equal numbers of males and females, a distribution across grades (with most at grades K–4), ethnic diversity, and diversity of socioeconomic status (education level of the examinee's parents).

Table 2 reports the characteristics of the sample of examinees. Overall, the sample met the demographic targets, although there was a preponderance of Hispanic examinees (7 of 10).

Five examiners conducted the subtest administrations, and nine individuals (including two of the examiners) served as scorers. All of these individuals were qualified and experienced in administering psychoeducational tests to children. The examiners and scorers received onsite training in administering and scoring the WIAT–III subtests with paper materials. The scorers conducted several practice scorings as well as a qualifying scoring activity that determined their ability to participate in the study. Examiners who were not Pearson employees were compensated for their participation.

Table 2 Demographic characteristics of the sample

Demographic Characteristic		*N*
Number of Cases		10
Grade	Kindergarten	2
	1	2
	2	1
	3	1
	4	2
	8	1
	9	1
Sex	Female	6
	Male	4
Ethnicity	African American	1
	Hispanic	7
	White	2
Parent Education	< 12 years	3
	HS graduate	1
	Some post–HS	3
	4–year degree	3

Procedure

Training and testing took place at Pearson's office in San Antonio, TX in May and June, 2013.

The two WIAT–III subtests were administered as part of a larger study that included three WISC–IV subtests (Block Design, Similarities, and Matrix Reasoning). The WISC–IV subtests were administered first (either in paper or Q-interactive format), and then the WIAT–III Oral Reading Fluency and Sentence Repetition subtests were administered, in that sequence, in paper format. To make the WIAT–III administrations as realistic as possible, the examiner performed all of the usual recording and scoring procedures; however, these scores were not used in the analysis. A video recording was made to capture the examiner's view of the administration, showing the examinee and the examinee tablet, but not showing the examiner's record form.

Each video-recorded administration was scored by nine scorers; four or five used the Q-interactive format and the remainder used the paper format. Each scorer scored all ten administrations, five in each format. To ensure that there was no correlation of formats among scorers, format assignments were made using a random number table to identify the scorers who would score each administration in Q-interactive format. A few adjustments were then made so that each scorer would do five scorings with each format.

Scoring was done independently by each scorer in an isolated room. The scorer watched the video of the administration, and recorded responses on the paper record form or the Q-interactive tablet. Scorers were encouraged to use any methods or Q-interactive features (such as audio capture) that they would use in clinical practice. They scored each trial in real time—they were not permitted to stop and restart the video during a trial, nor were they allowed to watch the video a second time. However, scorers were allowed to pause the video between trials.

As in the previous Q-interactive equivalence studies, video recordings were made of the scorings. These recordings served two purposes: first, in the event of a finding of non-equivalence, researchers could investigate possible causes by reviewing the scorers' behaviors; and second, the videos provided information about how scorers interacted with the digital and paper materials, which can be helpful in future test design. These videos were recorded from behind the scorer to show the monitor the scorer was watching and the digital tablet or record form on which the scorer was capturing responses and entering scores.

Pearson staff reviewed the scorer markups of text to ensure that the errors recorded were accurately totaled and the correct totals were entered (either on the paper record form or on the digital tablet). Because this transfer of error markings is a clerical procedure that happens post-administration, it is not considered an aspect of scoring related to the scoring format. However, the scorers' totals were retained for analysis.

The paper and Q-interactive mean scores were compared for the same set of administrations (examinees). The first step was to compute a mean paper-format score and a mean digital-format score for each administration, based on the four or five scorings of each administration using each format. Then an overall mean score for each format was calculated by averaging these within-administration means. This procedure gave equal weight to the paper and digital scoring formats.

The effect size for each score variable is the difference between paper-format and digital-format means divided by the population standard deviation of the score's normative metric (e.g., 15 for a WIAT–III standard score). A *z* test was applied to evaluate statistical significance, using the root mean square of the within-administration-and-format deviations from the mean as the standard error.

A positive value of the format effect indicates that the digital format yields higher scores than the paper format. The format effect is reported in standard score units, and the effect size expresses the format effect in standard deviation units.

Results

All 90 scorings of Sentence Repetition were usable (9 scorings x 10 administrations). One of the Oral Reading Fluency scorings in digital format was dropped from the study because the scorer neglected to start the timer on both trials; this left a total of 71 scorings available for analysis (8 or 9 scorings of each of 8 administrations).

There were twelve ORF scorings (eight in digital format and four in paper format) in which the examiner recorded an incorrect number of errors. These totals were corrected prior to analysis.

The magnitude and statistical significance of format effects are reported in Table 3. None of the format effects are statistically significant at the .05 level, and all effect sizes are 0.11 or smaller, well within the tolerance limits for the formats to be considered equivalent.

Table 3 Differences between standard scores obtained using paper and Q-interactive recording formats

Subtest and Score	Paper Scoring			Q-interactive Scoring			Difference	*z*	Effect size
	N	Mean	*SD*	*N*	Mean	*SD*			
Oral Reading Fluency									
Fluency	35	98.1	9.5	36	97.4	9.3	–0.70	–0.48	–0.05
Accuracy	35	98.8	11.5	36	100.4	12.4	1.60	0.40	0.11
Rate	35	98.1	9.1	36	97.4	8.7	–0.70	–0.56	–0.05
Sentence Repetition	45	97.6	17.4	45	96.4	16.4	–1.20	–0.34	–0.08

Note. *N* is the number of scorings; Mean and *SD* are based on the distribution of within-administration means.

Discussion

Both the preliminary study of readability and the primary study of the scorer interface indicate that equivalent results are obtained on the WIAT–III Oral Reading Fluency, Sentence Repetition, Word Reading, and Pseudoword Decoding subtests using Q-interactive and the standard paper format. These findings add to the body of evidence about the effects (or lack of effects) of features of interface design on examinee performance and the examiner's accuracy in recording and scoring. As this body of knowledge grows, it will support generalization to other tests of the same type and features.

Valuable information about using Q-interactive was obtained from viewing the video recordings of the scorings of Oral Reading Fluency. On three of the 72 digital scorings of passage readings, the scorer forgot to stop the timer when the examinee reached the end of the passage, causing the completion times to be overestimated by about 15 to 40 seconds. These incidents did not affect the overall study results, but they are a reminder that examiners need to be careful about starting and stopping the timer that is built in to the digital tablet. A second observation had to do with the eight instances in which the total number of errors entered into the counter on the digital tablet by the scorer was smaller than the actual number of errors the scorer had marked on the text. On some of these instances, the scorer evidently did not pay attention to the number displayed on the counter, leading to the discrepancy. Again, this points to the need for Q-interactive users to adjust to the new interface features.

References

Cohen, M. (1997). *Children's memory scale™.* Bloomington, MN: Pearson.

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 administered 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 administrations of cognitive tasks: Selected NEPSY®-II and CMS™ subtests. Q-interactive Technical Report 4.* Bloomington, MN: Pearson.

Delis, D., Kaplan, E., & Kramer, J. (2001). *Delis-Kaplan executive function system®.* Bloomington, MN: Pearson.

Delis, D., Kramer, J., Kaplan, E., & Ober, B. (2000). *California verbal learning test®,* second edition. Bloomington, MN: Pearson.

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY®*–second edition. Bloomington, MN: Pearson.

Wechsler, D. (2008). *Wechsler adult intelligence scales®–fourth edition.* Bloomington, MN: Pearson.

Wechsler, D. (2009). *Wechsler individual achievement test®–third edition.* Bloomington, MN: Pearson.

Wechsler, D. (2003). *Wechsler intelligence scales for children®–fourth edition.* Bloomington, MN: Pearson.