Equivalence of Q-interactive™ and Paper Administration of WMS®–IV Cognitive Tasks

Q-interactive Technical Report 6

Mark H. Daniel, PhD
Senior Scientist for Research Innovation

October 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 sixth Q-interactive equivalence study. In this study, the equivalence of scores from digitally assisted and standard administrations of the *Wechsler Memory Scale*®—fourth edition (WMS®–IV; Wechsler, 2009a) was evaluated.

In the first two equivalence studies, all fifteen *Wechsler Adult Intelligence Scale*®–fourth edition (WAIS®–IV; Wechsler, 2008) subtests and thirteen of fifteen *Wechsler Intelligence Scale 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 System*™ (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. Finally, the fifth study evaluated the Oral Reading Fluency and Sentence Repetition subtests of the *Wechsler Individual Achievement Test*®—third edition (WIAT®–III; Wechsler, 2009b), both of which met the equivalence criterion.

In all of 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 is 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. In each Q-interactive equivalence study, all or most of the 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, D-KEFS®, and WIAT®–III studies (Q-interactive Technical Reports 3 and 5). 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.

The current study of the WMS®–IV uses all three of these designs, because different subtests lend themselves to different methods. The retest and dual-capture methods require fewer cases than the equivalent-groups method, so they are used when possible. The equivalent-groups method is used when the characteristics of the subtest make the more efficient methods unsuitable.

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 slightly more than one-half of a point on the scaled-score metric used for WMS–IV subtests (mean of 10 and standard deviation of 3).

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.

WMS®–IV Equivalence Study

Measures

The WMS–IV is a comprehensive, individually administered measure of a variety of aspects of memory, including immediate and delayed recall and recognition of verbal and visual stimuli. All of the WMS–IV subtests were judged to require equivalence evaluation, because their Q-interactive interfaces have features that could plausibly affect how the examinee performs or how the examiner captures or scores responses, and these features have not previously been studied and are not already known to be free of format effects.

The WMS–IV subtests are described in Table 1. The examinee's digital tablet was used to present all visual stimuli. Examinees responded by touching the screen on Symbol Span, but examinees responded using paper for Clock Drawing and Visual Reproduction, and used a physical grid and cards for Designs and Spatial Addition.

Table 1 Description of WMS–IV subtests

Subtest	Description	Study Type
Brief Cognitive Status Exam™ (BCSE)	Variety of simple tasks: time orientation, mental control, clock drawing, incidental recall, auditory and inhibitory control, and verbal production	Retest and Dual Capture
Logical Memory	The examinee hears two short stories and retells each from memory.	Dual Capture
Verbal Paired Associates	The examinee hears a set of word pairs, and then produces the second word upon hearing the first.	Equivalent Groups
Designs	The examinee sees a grid with designs arrayed in the cells for 10 seconds, and then reproduces the array.	Retest
Visual Reproduction	The examinee sees an abstract design for 10 seconds, and then draws it on paper from memory.	Equivalent Groups
Spatial Addition	The examinee sees two grids with colored circles, one after the other, and then places circles in a third grid according to addition and subtraction rules.	Retest
Symbol Span	The examinee sees a series of abstract symbols, and then reproduces the series by selecting symbols in order.	Retest

As indicated in Table 1, different subtests lent themselves to different study designs. The dual-capture design is appropriate when the examinee does not interact directly with Q-interactive (i.e., the examinee tablet is not used) and when there is little if any interaction between examiner and examinee during administration of an item. Logical Memory and most tasks in the *Brief Cognitive Status Exam* (BCSE) met these criteria. A retest design is suitable when the examinee is unlikely to remember details of item responses and when learning a test-taking strategy is not a major determiner of performance. The retest method was used with the Designs, Spatial Addition, and Symbol Span subtests and the Inhibition portion of BCSE. The two remaining subtests, Visual Reproduction and Verbal Paired Associates, were studied using an equivalent-groups design. On subtests containing both an immediate trial (Part 1) and a delayed trial (Part 2), only the immediate trial was evaluated. The delayed trials did not require study because their examinee and examiner interfaces do not present any challenging features that were not included on the immediate trials.

For operational efficiency, the research was divided into two studies: Study A, which combined the retest and dual-capture procedures, and Study B, which used the equivalent-groups design.

Study A: Retest and Dual-Capture

Method

Participants

The examinee sample consisted of 30 individuals made up of 15 demographically matched pairs. All of these examinees were used in the retest analyses, and ten of them were also used for the dual-capture analyses.

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; 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 ages, ethnic diversity, and diversity of socioeconomic status (education level of the examinee or, at ages 16–24, the examinee's parents).

Table 2 reports the characteristics of the sample. Both the retest and dual-capture samples had a good distribution of demographic characteristics. In the retest sample, examinees within each pair were matched precisely on age group, sex, and ethnicity. Education level was precisely matched in nine of the pairs, but differed by one level in four of the other pairs and differed by two levels the other pair.

Eight examiners conducted the retest administrations. All of these individuals were qualified and experienced in administering psychological tests. The examiners received onsite training in administering and scoring the subtests with paper materials, and conducted several practice administrations as well as a qualifying administration. They also received training in using Q-interactive to administer and score the subtests. Examiners who were not Pearson employees were compensated for their participation.

The ten administrations used in the dual-capture portion of the study were conducted by four of the examiners. There also were 10 individuals (including two of the examiners) who scored the video recordings of these administrations. These scorers had experience and study-specific training, including practice scorings, comparable to the examiners.

.

Table 2 Study A Retest and Dual Capture: Demographic characteristics of the sample

Demographic Characteristic		Retest: Paper–Digital	Retest: Digital–Paper	Dual Capture
Total		15	15	10
Age	16–24	4	4	3
	25–54	8	8	4
	55–69	3	3	3
Sex	Female	9	9	6
	Male	6	6	4
Ethnicity	African American	1	1	1
	Hispanic	6	6	3
	White	7	7	6
	Other	1	1	0
Education[a]	< 12 years	3	1	0
	HS graduate	3	5	3
	Some post-HS	4	4	3
	4–year degree	4	5	4

[a] Education level was not available for one examinee in the paper–digital retest sequence group.

Procedure

Each examinee took the following sequence of subtests:

- Format 1 (for retest analysis)
  - Designs
  - Spatial Addition
  - BCSE Inhibition task
  - Symbol Span
- BCSE non-Inhibition tasks (paper format, for dual-capture analysis)
- Logical Memory (paper format, for dual-capture analysis)
- Format 2 (for retest analysis)
  - Designs
  - Spatial Addition
  - BCSE Inhibition task
  - Symbol Span

Examinees were not told in advance that they would be taking some of the subtests a second time. All examinees took the BCSE non-Inhibition tasks and Logical Memory, regardless of whether they were participating in the dual-capture study, to make the test–retest interval consistent for all examinees. For half of the examinees (one member of each matched pair), Format 1 was paper and Format 2 was Q-interactive; for the other half, the formats were reversed.

Examiners performed all of the usual recording and scoring procedures, including post-administration scoring of examinee responses. Examiners' scores were used in the retest analysis, but were not used in the dual-capture analysis. In all instances of paper administration, Pearson staff reviewed the conversion of item scores to subtest raw scores to eliminate any computation errors. (This step is performed automatically by Q-interactive.) Because calculating subtest raw scores is a clerical procedure that happens post-administration, it is not considered to be related to the administration format.

For the dual-capture analysis, the administrations of BCSE (non-Inhibition) and Logical Memory were video recorded to capture the examiner's view of the administration, showing the examinee, the stimulus book, and the response booklet, but not showing the examiner's record form. Each video-recorded administration was scored a total of 10 times by 10 different scorers. Half of the scorings of each administration were performed using 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 between formats and scorers, a random number table was used 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 subtest or task in real time—they were not permitted to stop and restart the video during a subtest or task, nor were they allowed to watch the video a second time. Scorers were allowed to pause the video only between subtests or tasks.

As in the previous Q-interactive equivalence studies, video recordings were made of the retest administrations and the dual-capture scorings. These recordings served two purposes: first, in the event of a finding of non-equivalence, researchers could investigate possible causes by reviewing the behavior of examinees, examiners, and scorers; and second, the videos provided information about how individuals interacted with the digital and paper materials, which can be helpful in future test design. The videos of dual-capture scoring 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.

Analysis

Retest. The retest analysis focused on the "change score" for each examinee (i.e., the change in score from the first administration to the second administration). If there was no effect of format, the average change score would be the same for the paper–digital and digital–paper sequence groups (except for sampling error and measurement error). If there was a format effect, the average

change scores in the two sequence groups would differ by twice the size of the effect, because in one sequence group the effect would increase the average difference score and in the other sequence group it would reduce it. The format effect was calculated by subtracting the average change score in the digital–paper sequence from the average change score in the paper–digital sequence, and dividing by two. A positive value indicated that the digital format yielded higher scores than the paper format. As in the dual-capture analysis, the format effect was expressed in the subtest's score units, and the effect size expressed the format effect in standard deviation units.

***Dual Capture*.** The analysis compared the mean scores from paper and Q-interactive scorings of 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 five scorings of each administration using each format. Then an overall mean score for each format was calculated by averaging these within-administration means across administrations. This procedure gave equal weight to the paper and digital scoring formats.

The effect size for each score was the difference between paper-format and digital-format means divided by the population standard deviation of scores on that subtest. For Designs, Spatial Addition, and Symbol Span, which are scored using scaled scores, the standard deviation was 3. For each of the subscores of Inhibition, which reports "weighted scores," the standard deviation was the root-mean-square average of within-age-group standard deviations in the WMS-IV norm sample, weighted to reflect the age distribution in this study. 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.

Regardless of the study design, a positive value of the format effect and effect size indicated that the digital format yielded higher scores than the paper format.

Results

***Retest*.** One of the BCSE Inhibition administrations was not usable (for reasons unrelated to the administration format), so the analysis of those tasks was based on 14 of the 15 matched pairs of examinees. The post-administration calculation of Designs raw scores by Pearson staff identified and corrected a considerable number of calculation errors that had been made by examiners. Table 3 reports average scaled scores (or weighted scores for Inhibition) for each administration format within each administration-sequence group. On the subtests that use scaled scores, the subgroup taking Q-interactive administrations first had first-trial means that were very close to the population values. In the subgroup taking the paper format first, first-trial means were variable and the standard deviations were consistently smaller than population values. In both subgroups, the practice effect on the Spatial Addition subtest was smaller than on Designs or Symbol Span.

Table 3 Study A Retest: Mean (*SD*) scores[a] by sequence and administration format

		Digital–Paper				Paper–Digital			
Subtest	**Pairs**	**Trial 1**		**Trial 2**		**Trial 1**		**Trial 2**	
Designs	15	9.80	(3.17)	12.87	(2.75)	7.87	(2.33)	10.47	(2.92)
Spatial Addition	15	9.93	(3.63)	10.07	(3.31)	11.07	(2.58)	11.87	(2.70)
Symbol Span	15	9.87	(2.03)	12.00	(2.10)	8.67	(2.13)	10.47	(1.88)
BCSE Inhibition									
Time	14	3.29	(1.07)	3.71	(0.47)	3.60	(0.74)	4.00	(0.00)
Omission Errors	14	3.71	(1.07)	4.00	(0.00)	3.47	(1.41)	3.71	(1.07)
Commission Errors	14	6.79	(2.89)	7.36	(0.84)	6.53	(2.61)	8.00	(0.00)

[a] Scaled scores for Designs, Spatial Addition, and Symbol Span; weighted scores for Inhibition tasks

Table 4 shows the results of the analysis of format effects. None of the effects was statistically significant. Effect sizes for the three complete subtests (Designs, Spatial Addition, and Symbol Span) were small (ranging from –0.06 to 0.11) and well within the range of +/– 0.20 that is the criterion for equivalence. Two of the subtasks of the Inhibition section of the *Brief Cognitive Status Exam* showed no format effect, but the third (Commission Errors) had an effect size of 0.35, exceeding the equivalence criterion. All video recordings and record forms were reviewed to verify that there were no administration errors or other anomalies affecting Inhibition.

Table 4 Study A Retest: Format effects

	Change Score						
	Digital–Paper		Paper–Digital		Format		Effect
Subtest	**Mean**	***SD***	**Mean**	***SD***	**Effect**	***t***	**Size**
Designs	3.07	1.87	2.60	2.87	–0.24	–0.48	–0.08
Spatial Addition	0.13	1.51	0.80	2.04	0.34	1.13	0.11
Symbol Span	2.13	1.92	1.80	1.15	–0.17	–0.54	–0.06
BCSE Inhibition							
Time	0.43	0.76	0.43	0.76	0.00	0.00	0.00
Omission Errors	0.29	1.07	0.29	1.90	0.00	0.00	0.00
Commission Errors	0.57	3.03	1.57	2.68	0.50	0.89	0.35

Dual Capture. Five scorings were available in each format for each administration, except for the paper scoring of the BCSE Clock Drawing task where only four scorings could be used. Average scores for each format, and format effects, are reported in Table 5. One task within BCSE (Clock Drawing) showed an effect size that exceeded the criterion of 0.20, although it was not statistically significant; on this task, scores were lower using the digital scoring interface. Close inspection of the scorings did not reveal any systematic errors or other explanations of this effect. However, format had only a negligible effect on the total BCSE score, with an effect size of –0.06.

Table 5 Study A Dual Capture: Differences between scores[a] obtained using paper and Q-interactive recording formats

Subtest	Paper Scoring		Q-interactive Scoring		Format Effect	*z*	*SD* (norm)	Effect Size
	Mean	*SD*	Mean	*SD*				
Orientation	6.50	1.58	6.50	1.58	0.00	0.00	1.36	0.00
Time Estimation	2.62	1.36	2.50	1.24	–0.12	–1.00	0.80	–0.15
Mental Control: Time	3.38	0.95	3.32	0.95	–0.06	–1.00	0.76	–0.08
Mental Control: Errors	5.51	2.45	5.44	2.40	–0.07	–0.30	2.09	–0.03
Clock Drawing	3.88	0.21	3.70	0.53	–0.18	–1.87	0.56	–0.32
Incidental Recall	7.60	1.26	7.44	1.37	–0.16	–1.00	0.97	–0.17
Verbal Production	5.80	0.51	5.63	0.63	–0.17	–2.68**	0.92	–0.19
BCSE Total					–0.26		4.29	–0.06
Logical Memory	10.48	2.73	10.30	2.71	–0.18	–0.91	(3)	–0.06

** $p < .01$

[a] Weighted scores for BCSE; scaled scores for Logical Memory

Note. The BCSE Total is the sum of weighted scores for all tasks, including Inhibition. $N = 50$ scorings in each format except for paper scoring of Clock Drawing, for which $N = 49$. Mean and *SD* are based on the distribution of within-administration means.

Study B: Equivalent Groups

Method

Participants

A total of 99 examinees took the Visual Reproduction and Verbal Paired Associates subtests in the Q-interactive format, along with the covariate tests in paper format. The paper-administration comparison group consisted of the 186 members of the WMS–IV norm sample who had also taken the CVLT-II. For the Q-interactive sample, the inclusion requirements, demographic requirements, and recruiting procedures were similar to those for Study A. None of these examinees had a clinical diagnosis or special-education classification, although three had a designation of gifted/talented. The demographic characteristics of the Q-interactive sample are presented in Table 6.

Table 6 Study B: Demographic characteristics of the Q-interactive sample

Demographic Characteristic		Examinees
Total		99
Age	16–24	24
	25–43	48
	55–69	27
Sex	Female	54
	Male	45
Ethnicity	African American	12
	Asian	3
	Hispanic	21
	White	59
	Other	4
Education	< 12 years	4
	HS graduate	27
	Some post-HS	33
	4-year degree	35

Procedure

All examinees in the Q-interactive sample took the subtests in the following sequence:

- WMS–IV Visual Reproduction Part 1 (Q-interactive format)
- WMS–IV Verbal Paired Associates Part 1 (Q-interactive format)
- WMS–IV Designs Part 1 (paper format)
- CVLT-II Immediate and Short Delay trials (paper format)

Examiners were qualified and experienced in administering psychological tests. They received training in administering WMS–IV with Q-interactive and they conducted several practice administrations before the study began. Testing took place in San Antonio, Texas; Detroit, Michigan; and Fort Collins, Colorado. Almost all of the administrations were video recorded so that the accuracy of both digital and paper administrations could be evaluated if any format effects were found, and to provide information on how examinees and examiners interacted with the Q-interactive interface. All examinees and examiners who were not Pearson employees, were paid for their participation.

Examiners performed all of the usual recording and scoring procedures that are required during item administration. In addition, they used the Q-interactive examiner interface to score the Visual Reproduction responses post-administration. Total scores on the paper-administered tests (Designs and CVLT-II) were calculated by Pearson staff using the responses and scores recorded in the record form by the examiners.

Analysis

The 186 WMS–IV norm sample cases were used to develop multiple regression equations in which the Designs subtest scaled score, several CVLT-II scores (Free-Recall Trial A *t* score, Free-Recall Trial B *z* score, and Short-Term Free-Recall and Cued Recall *z* scores), and demographic characteristics (age, sex, ethnicity, and education) were the independent predictors and the scaled score for either Visual Reproduction or Verbal Paired Associates was the dependent variable. Then, the actual Visual Reproduction and Verbal Paired Associates scaled scores for the Q-interactive sample were compared with scores predicted from these regression equations. The residuals (observed minus predicted scores) represent the digital effect plus error. The analysis of each subtest used a one-sample *t* test with the null hypothesis stating that the average residual is zero.

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., 3 for a WMS–IV scaled score). A *t* test was applied to evaluate statistical significance of the hypothesis that the difference between means was not equal to 0. 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 scaled score units and the effect size expresses the format effect in standard deviation units.

Results

Four Visual Reproduction cases could not be used because of errors in administration that were unrelated to the Q-interactive format. Table 7 reports descriptive statistics for the four subtests in the Q-interactive sample. Means and standard deviations of WMS–IV subtest scaled scores were near the population averages of 10 and 3, and the CVLT-II scores, likewise, were close to their corresponding population values, indicating that the sample is reasonably representative of the population.

The multiple correlations of covariate test scores and demographics with Visual Reproduction and Verbal Paired Associates scaled scores were .58 and .61, respectively, in the WMS–IV norm sample and .46 and .63 in the Q-interactive sample.

The magnitude and statistical significance of format effects are reported in Table 7. The Q-interactive effect sizes for Visual Reproduction (–0.13) and Verbal Paired Associates (–0.03) were well within the equivalence standard of 0.20 and were not statistically significant.

Table 7 Study B: Descriptive statistics and effect sizes

Test/Subtest	*N*	Mean	*SD*	Residual[a] Mean	Residual[a] *SD*	*t*	Effect Size
Paper-administered covariate tests:							
Designs (scaled score)	99	9.3	2.6				
CVLT-II[b]							
Free Recall List A (*t*)	99	55.9	11.3				
Free Recall List B (*z*)	99	–0.2	1.0				
Short-Delay Free Recall (*z*)	99	0.3	1.2				
Short-Delay Cued Recall (*z*)	99	0.1	1.0				
Q-interactive-administered subtests:							
Visual Reproduction (scaled score)	94	9.9	2.9	–0.40	2.56	–1.55	–0.13
Verbal Paired Associates (scaled score)	99	10.1	2.7	–0.09	2.07	–0.45	–0.03

[a] Actual score minus predicted score

[b] *t*-score mean = 50, *SD* = 10; z-score mean = 0, *SD* = 1

Discussion

None of the WMS–IV subtests showed an effect of the Q-interactive format that reached an effect size value of 0.20, indicating that Q-interactive administration yields equivalent results as paper administration. Two tasks within the optional *Brief Cognitive Status Exam*, Clock Drawing and Inhibition Commission Errors, showed effect sizes of –.32 and .35, respectively. However, these effects were not statistically significant, and the overall BCSE score had a negligible effect size of –0.06. These findings add to the body of research indicating that when tests originally designed for paper-based administration are carefully adapted for administration on the Q-interactive platform, results are comparable using either administration format.

References

Cohen, M. (1997). *Children's memory scale™.* 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. (2003). *Wechsler intelligence scale for children®*–fourth edition. Bloomington, MN: Pearson.

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

Wechsler, D. (2009a). *Wechsler memory scale®*–fourth edition. Bloomington, MN: Pearson.

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