Equivalence of Q-interactive™ and Paper Administrations of Cognitive Tasks: Selected NEPSY®–II and CMS Subtests

Q-interactive Technical Report 4

Mark H. Daniel, PhD
Senior Scientist for Research Innovation

May 2013

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 formats. If equivalence is demonstrated, then the existing norms, reliability, and validity information can be applied to Q-interactive results.

This is the fourth Q-interactive equivalence study. In this study, the equivalence of scores from digitally assisted and standard administrations of three NEPSY®–II (Korkman, Kirk, & Kemp, 2007) subtests (Memory for Designs—Immediate, Picture Puzzles, and Inhibition) and two *Children's Memory Scale* (CMS; Cohen, 1997) subtests (Picture Locations and Dot Locations) were evaluated.

In the first two equivalence studies, all fifteen *Wechsler Adult Intelligence Scales*®–fourth edition (WAIS®–IV) subtests and thirteen of fifteen *Wechsler Intelligence Scales for Children*®–fourth edition (WISC®–IV) 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 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. One might say that 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 this 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 in this way 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 had the additional benefit of showing how examiners and examinees interacted with the test materials in each format.

As a whole, the equivalence studies indicate that examinees age 6 and older (the youngest individuals tested) 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 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.

Another type of 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.

A third design, *retest,* was used in the follow-up study of the WAIS–IV Processing Speed subtests (Technical Report 1) and is used in this study of selected NEPSY–II and CMS subtests. 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. There was judged to be little risk of a change in cognitive process in the second administration of the Processing Speed tasks. (This judgment was supported by the fact that the findings were the same as in the initial equivalent-groups study.) The current study involves those subtests on NEPSY–II and CMS that were judged to require an equivalency study but to have little risk of changes in cognitive process on retest.

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 scaled-score point on the commonly used subtest metric that has a 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 with a new format takes a substantial amount of practice.

NEPSY–II and CMS Equivalence Studies

Measures

The three NEPSY–II subtests and two CMS subtests were chosen for this study because they met two criteria:

- Their Q-interactive examinee and/or examiner interfaces have features that could plausibly affect examinee performance or the examiner's ability to administer, capture, and score accurately, but which have not previously been studied and are not already known to be free of effects on equivalence.
- The cognitive processes used during a second administration were judged unlikely to be significantly affected by the examinee's having taking the subtest a short time previously. Three of the subtests involve short-term memory for non-meaningful visual information; one involves visual matching; and the fifth requires the examinee to follow rules that govern his or her response to visual stimuli. Specific item content would be very difficult to remember, and these tasks do not lend themselves to problem-solving strategies.

Though the three memory subtests have delayed trials, only the immediate trials are included in this study. The delayed trials do not introduce any new interface features.

NEPSY–II Subtests

The NEPSY–II is a comprehensive instrument used to assess neuropsychological development in preschool and school-age children. The three subtests included in this study are:

Inhibition. This subtest measures the ability to inhibit automatic responses and switch between response types. The child sees rows of black and white shapes or arrows. On the Naming trial, the child must say the actual shape (or direction) of each stimulus. On the Inhibition trial, the child must say the other shape (or the opposite direction) of each stimulus. Finally, on the Switching trial (administered at ages 7 and older), the child must say the actual shape (or direction) if the stimulus is black and the other shape (or the opposite direction) if the stimulus is white. Scoring is based on errors and time. In the Q-interactive version, the examinee screen displays the rows of stimuli (arranged just as in the stimulus book), and the examiner captures responses by touching the stimuli arrayed on the examiner screen.

Memory for Designs. This subtest measures spatial memory for novel visual material. The child sees a grid with abstract designs in some of the cells, for 10 seconds. Then the child places cards containing those designs in the correct locations in a blank grid. In the Q-interactive version, the examinee screen displays the stimulus grid, but the child responds using a physical grid and physical cards, just as in the paper version.

Picture Puzzles. This is a measure of visual discrimination, spatial localization, and visual scanning. The child sees a large picture divided into a grid and four smaller pictures that match cells of the grid. The child must match the smaller pictures to the cells in the grid from which each was taken. In the Q-interactive version, the examinee screen displays the pictures and captures the child's touch response. Picture Puzzles is administered at ages 7 and older.

CMS Subtests

The *Children's Memory Scale* consists of subtests used to measure diverse aspects of verbal and visual memory. The two subtests included in this study are:

Dot Locations. This subtest measures memory for spatial location. The procedure and materials are similar to those for NEPSY–II Memory for Designs, except that the stimuli in the grid cells are circles, and the exposure time is five seconds. The stimulus grid is shown on the examinee screen, and the examinee responds using physical chips and a physical grid.

Picture Locations. This subtest also measures memory for spatial location. The procedure and materials are the same as for Dot Locations, except that the stimuli are illustrations of objects (e.g., train, cat, or car), and the exposure time is two seconds. All of the stimuli on an item are of the same object.

Method

Participants

The sample consists of 20 demographically matched pairs of 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 severe clinical conditions. The sampling plan called for approximately equal numbers of males and females, a wide distribution of ages, ethnic diversity (at least three pairs each representing African American, Hispanic, and White groups), and diversity of socioeconomic status (education level of the examinee's parents). Pairs of examinees were matched by age range, gender, ethnicity, and parent education. (In two of the twenty pairs, the parent education level differed by 1 between the members of the pair.) The first member of each pair was randomly assigned to one of the administration sequences (paper–digital or digital–paper), and the other member of that pair was assigned to the other sequence.

Because NEPSY–II Picture Puzzles and the Switching trial of Inhibition are not administered until age 7, the samples for those measures consisted of 15 rather than 20 matched pairs.

Table 1 reports the characteristics of the sample that took each sequence. Overall, the sample closely reflects the U.S. child population on gender and ethnicity, and there is a slight overrepresentation of higher levels of parent education (65%–70% with post–secondary education, compared with approximately 60% in the population).

The thirteen examiners were qualified and experienced in administering psychological and neuropsychological tests to children and adults. Seven of them had participated in a previous Q–interactive equivalence study. The examiners received onsite training in administering the NEPSY–II and CMS subtests both with paper materials and with Q–interactive. They conducted several practice administrations as well as a qualifying administration that determined their ability to participate in the study. Examiners who were not Pearson employees were compensated for their participation.

Table 1 Demographic characteristics of the sample by sequence group (*n* = 20 matched pairs)

Demographic Characteristic		Digital–Paper	Paper–Digital
Number of Cases		20	20
Age (years)	5–6	5	5
	7–9	5	5
	10–12	4	4
	13–16	6	6
Sex	Female	10	10
	Male	10	10
Ethnicity	African American	3	3
	Hispanic	5	5
	White	11	11
	Other	1	1
Parent Education	< 12 years	1	1
	HS graduate	5	6
	Some post–HS	2	2
	4–year degree	12	11

Procedure

Training and testing took place at Pearson's office in San Antonio, TX between December 15, 2012 and January 29, 2013.

The five subtests were administered in the following sequence:

- NEPSY–II Memory for Designs
- NEPSY–II Inhibition
- CMS Dot Locations
- NEPSY–II Picture Puzzles
- CMS Picture Locations

Each examinee took all five subtests in one format and then took them again in the other format, in the same test session. Examinees were not told at the beginning that they would be taking the subtests a second time. During paper administrations, examiners captured response information in the standard manner using a paper record form. The Pearson research team performed the post–administration scoring. The scoring process is not directly affected by the administration format.

As in the previous Q–interactive equivalence studies, administrations were video–recorded. These recordings served two purposes. First, in the event of a finding of non–equivalence, they enabled the researchers to investigate possible causes by reviewing the behavior of the examiners and examinees. Second, they provided information about how examiners interact with the digital and paper materials, which can be helpful in future test design. The videos were shot from behind and above the examiner and showed the examiner's and examinee's tablets and hands, and the physical test materials.

The analysis of a retest equivalence study focuses on the "change score" for each examinee (i.e., the change in score from the first administration to the second administration). If there is no effect of format, the average change score will be the same for the paper–digital and digital–paper sequence groups (except for sampling error and measurement error). If there is a format effect, the average change scores in the two sequence groups will differ by twice the size of the effect, because in one sequence group the effect will increase the average difference score and in the other sequence group it will reduce it. The format effect is 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 indicates that the digital format yields higher scores than the paper format. The format effect is expressed in scaled-score units. The effect size expresses the format effect in standard–deviation units (3 in the case of scaled scores).

Using demographically matched pairs of examinees in the two sequence groups produces high statistical power with small sample sizes. Assuming a retest correlation of 0.8, a sample of 15 matched pairs is needed to achieve power of 0.8 to detect an effect size of 0.2 (alpha = .05).

Results

The number of matched pairs that could be analyzed was slightly fewer than 20 for some measures because of administration errors unrelated to the Q-interactive format. Whenever such an error occurred on one case, the corresponding score on the matching case was also deleted from the analysis. (Picture Puzzles and Inhibition Switching Time have a maximum of 15 cases because they are administered at ages 7 and older.)

Table 2 reports the number of matched pairs for each measure, and the means and standard deviations of scaled scores for the first and second administrations in each sequence group (paper–digital and digital–paper). In almost all instances, scores were higher on the second administration than the first.

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 0.20 or smaller, so they are within the tolerance limits for considering the formats to be equivalent. Most of the format effects are negative, indicating a tendency for scores to be slightly higher with the paper format than the digital format.

Table 2 Means *(SD)* of scaled scores, by sequence group and administration format

Subtest	Pairs	Digital–Paper		Paper–Digital	
		Trial 1	Trial 2	Trial 1	Trial 2
NEPSY–II					
Inhibition					
Naming Time	19	8.7 (2.7)	9.7 (3.2)	8.8 (2.4)	9.3 (2.6)
Inhibition Time	20	10.1 (2.7)	11.2 (3.6)	8.8 (2.7)	9.6 (2.1)
Switching Time	11	10.1 (2.0)	12.6 (2.6)	10.2 (1.7)	11.4 (2.5)
Errors	16	9.4 (2.9)	11.2 (1.9)	7.9 (2.8)	9.6 (4.5)
Memory for Designs					
Content	18	10.1 (2.7)	13.0 (1.6)	9.2 (3.7)	11.4 (2.9)
Spatial	18	9.9 (2.7)	12.0 (1.8)	9.9 (3.3)	10.8 (2.5)
Total	18	9.7 (2.5)	13.4 (1.8)	9.1 (4.0)	11.8 (2.8)
Picture Puzzles	15	9.9 (3.6)	11.0 (3.4)	9.7 (4.0)	11.9 (3.0)
CMS					
Dot Locations	19	10.9 (2.6)	13.1 (2.7)	10.6 (3.0)	11.8 (2.8)
Picture Locations	20	11.0 (3.3)	11.6 (2.9)	10.4 (2.8)	10.2 (3.3)

Table 3 Format effects: Differences between scores obtained using paper and Q–interactive administration formats

Subtest	Change Score				Format Effect	*t*	Effect Size
	Digital–Paper		Paper–Digital				
	Mean	*SD*	Mean	*SD*			
NEPSY–II							
Inhibition							
Naming Time	0.9	2.7	0.5	2.6	–0.21	–0.49	–0.07
Inhibition Time	1.2	2.5	0.8	3.1	–0.18	–0.39	–0.06
Switching Time	2.7	2.1	2.0	0.9	–0.36	–1.08	–0.12
Errors	1.8	3.0	2.8	3.8	0.50	0.82	0.17
Memory for Designs							
Content	2.9	2.2	2.3	2.8	–0.33	–0.80	–0.11
Spatial	2.1	2.7	0.9	3.1	–0.61	–1.27	–0.20
Total	3.8	2.4	2.7	3.0	–0.56	–1.24	–0.19
Picture Puzzles	1.1	1.9	2.2	2.6	0.57	1.35	0.19
CMS							
Dot Locations	2.1	2.3	1.3	1.5	–0.50	–1.50	–0.17
Picture Locations	0.6	3.0	–0.2	2.5	–0.40	–0.93	–0.13

Note. See text for definition of format effect. A positive format effect indicates higher scores on digital administration. Effect size = format effect / 3

Discussion

The five subtests investigated in this study contain digital interface features that had not previously been evaluated for possible effects on equivalence. Each of the three memory subtests (Memory for Designs, Dot Locations, and Picture Locations) displays a visual stimulus (grid) and requires the examinee to respond by placing cards or chips in a physical grid. Equivalence might be affected by displaying the stimulus on a digital tablet rather than a printed stimulus book, or by the difference in the way the examiner records responses on the tablet rather than in a printed record form. Results indicated that the format did not affect scores.

The Picture Puzzles subtest was studied because the examinee must carefully analyze a complex visual stimulus shown on the tablet, which is the type of interface that showed a small format effect (favoring Q–interactive) in the WISC–IV study. In this study, the format effect was nonsignificant and smaller than the 0.20 threshold for effect size, although it was in the same direction as the WISC–IV effects.

Finally, the study of the Inhibition subtest demonstrated that the relatively complex examiner interface required to capture the examinee's rapid oral responses functioned effectively and had no effect on score levels. This finding is consistent with that for the D-KEFS Color–Word Interference subtest which uses a somewhat similar examiner interface for response capture. Unlike Color–Word Interference, Inhibition used the tablet to display the visual stimulus, but the stimulus is simple rather than detailed (rows of simple black or white shapes), so the lack of a positive format effect (i.e., favoring digital presentation) is not surprising.

As a consequence of the highly efficient nature of the retest study design, there were not enough cases in this study to permit an evaluation of the influence of demographic characteristics (age, gender, ethnicity, or socioeconomic status) on format effects.

These NEPSY–II and CMS equivalence studies add to the body of evidence about the effect (or lack of effect) of features of interface design on how examiners capture and score responses. As this body of knowledge grows, it will support generalization to other tests of the same type and 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.

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