Equivalence of Q-interactive® and Paper Administrations of Language Tasks: Selected CELF®–5 Tests

Q-interactive Technical Report 7

Mark H. Daniel, PhD

Dustin Wahlstrom, PhD

Xuechun Zhou, PhD

June 2014

Pearson

Introduction

Q-interactive®, a Pearson digital system 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 system, the goal has been to maintain raw-score equivalence between standard (paper) and digital administration and scoring formats. If equivalence is demonstrated, then the norms, reliability, and validity information gathered for the paper format can be applied to Q-interactive results.

This is the seventh Q-interactive equivalence study. In this study, the equivalence of scores from digitally assisted and standard administrations of four tests of the *Clinical Evaluation of Language Fundamentals*®–fifth edition (CELF®–5; Wiig, Semel, & Secord, 2013) were evaluated. (Note that CELF–5 refers to each task as a *test* rather than a *subtest*.)

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 Scale*™ (D-KEFS™; Delis, Kaplan, & Kramer, 2001) subtests and the Free-Recall trials of the *California Verbal Learning Test*®–second edition (CVLT®–II; Delis, Kramer, Kaplan, & Ober, 2000), 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. The fifth study evaluated the Oral Reading Fluency and Sentence Repetition subtests of the *Wechsler Individual Achievement Test*®—third edition (WIAT®–III; Wechsler, 2009a), both of which met the equivalence criterion. In the most recent study, all subtests of the *Wechsler Memory Scale*®–fourth edition (WMS®–IV; Wechsler, 2009b) 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. Although 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.* To date, most of the differences between paper and Q-interactive administrations have 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 has been 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.

- *Q-interactive system scoring the examinee's touch responses accurately.* With CELF–5, Q-interactive introduces automatic scoring. Previous implementations have required the examiner to enter a score for each item, which maintains examiner control, but does not take advantage of the capabilities of the tablet system to recognize and score touch responses. The dependability of this technology needed to be evaluated, particularly in the case of items requiring a pattern or sequence of touches.

- *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 was 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 5 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 uncovered so far (WISC–IV 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*, 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. Details about this design may be found in Technical Reports 3 (CVLT–II and D-KEFS), 5 (WIAT–III), and 6 (WMS–IV).

In the third design, *retest,* 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 retest design has been used in the follow-up study of WAIS–IV Processing Speed subtests (Technical Report 1) and in the studies of the NEPSY–II (Technical Report 4) and WMS–IV (Technical Report 6); it is used in the present study of CELF–5.

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 tests' 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.

CELF–5 Equivalence Study

Method

Measures

CELF–5 is a comprehensive instrument used to assess a variety of expressive and receptive language skills at ages 5 to 21. Four CELF–5 tests were identified for study by the research team because their Q-interactive examinee and/or examiner interfaces have features that

- could plausibly affect the examiner's ability to administer, capture, and score accurately, or that could affect the accuracy of automatic scoring, and
- differ from the test formats already shown to be equivalent to a paper-and-pencil administration in other studies.

These tests are:

- Linguistic Concepts (ages 5–8): the examinee hears a spoken direction involving basic concepts (such as *and* or *before*), and responds by touching one or more pictures on the tablet, sometimes in a specified sequence
- Recalling Sentences (ages 5–21): the examinee hears a sentence and repeats it; the examiner transcribes what the examinee says by editing the printed sentence.
- Following Directions (ages 5–21): the examinee hears spoken directions of increasing length and complexity, and responds by touching one or more pictures on the tablet, in a particular sequence.
- Formulated Sentences (ages 5–21): the examinee must construct and say a sentence that uses particular word(s) and is consistent with an illustration; the examiner transcribes what the examinee says.

Recalling Sentences and Formulated Sentences are included in the study because they require the examiner to transcribe the examinee's oral response accurately and completely, so that it can be scored later if necessary. Linguistic Concepts and Following Directions were chosen because some of their items ask for a pattern or sequence of touch responses, which presents a challenge for automatic scoring by the Q-interactive system.

Participants

The target sample consisted of 20 demographically matched pairs of examinees, of which 14 were in the age range (5 to 8) at which Linguistic Concepts is administered. 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 distribution of ages, ethnic diversity, 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. 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.

The seven examiners were qualified and experienced in administering speech/language tests to children and adults. The examiners received onsite training in administering the CELF–5 tests 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.

Procedure

On the four CELF–5 tests, the cognitive processes used during a second administration were judged unlikely to be significantly affected by the examinee's having taking the test a short time previously. Specific item content would be difficult to remember, and these tasks do not lend themselves to problem-solving strategies. For these reasons, the retest design was selected. The Formulated Sentences and Recalling Sentences tests could have been studied using the dual-capture method, but because only four tests were being evaluated, it was judged to be most efficient to use the retest design for all of them.

Training and testing took place at Pearson's office in San Antonio, TX between February and April, 2014. The tests were administered in the following sequence:

- Linguistic Concepts (ages 5–8 only)
- Recalling Sentences
- Following Directions
- Formulated Sentences

Each examinee took the three or four age-appropriate tests 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 tests a second time. During paper administrations, examiners captured response information in the standard manner using a paper record form, and scored each item. The Pearson research team checked cases for proper use of administration rules (such as start points and discontinue), but did not rescore items.

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 would enable the researchers to investigate possible causes by reviewing the behavior of the examiners and examinees. Second, they would provide information about how examiners and examinees interact with the digital and paper materials, which can be helpful in future test design. The videos were shot from above and to the side of the examiner and examinee and showed both tablets.

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 expected change score will be the same for the paper–digital and digital–paper sequence groups. If there is a format effect, the change scores in the two sequence groups will be expected to 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 the format effect 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 2. 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).

Statistical significance of the format effect is calculated using a paired-samples *t* test: the change score for an examinee in the digital–paper group is subtracted from the change score for the matched examinee in the paper–digital group, and the mean and standard deviation of these differences generate the *t* score. 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

Table 1 reports the characteristics of the sample that took each sequence. The number of matched pairs that could be analyzed was 18 (out of the target of 20) for Following Directions and

Formulated Sentences, 17 (out of 20) for Recalling Sentences, and 13 (out of 14) for Linguistic Concepts because of administration errors unrelated to the Q-interactive format. Whenever such an error occurred, the score for that test on the other administration format for that examinee, and both scores on the matching case, were also deleted from the analysis.

Table 1 Demographic characteristics of the sample by sequence group

		RS, FD, and FS		Linguistic Concepts	
Demographic Characteristic		**Digital–Paper**	**Paper–Digital**	**Digital–Paper**	**Paper–Digital**
Number of Cases		18	18	13	13
Age (years)	5–6	6	7	6	7
	7–8	8	7	7	6
	9–13	4	4	0	0
Sex	Female	9	9	6	6
	Male	9	9	7	7
Ethnicity	African American	4	4	3	3
	Hispanic	5	5	3	3
	White	8	8	6	6
	Other	1	1	1	1
Parent Education	< 12 years	1	4	1	3
	HS graduate	6	2	3	2
	Some post-HS	3	6	1	2
	4-year degree	8	6	8	6

The digital–paper and paper–digital samples are perfectly matched on sex and ethnicity, are nearly identical on age, and are similar on parent education. Overall, the sample closely reflects the U.S. child population on sex, ethnicity, and parent education

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). Scores usually were higher on the second administration than the first.

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

		Digital–Paper Sequence		Paper–Digital Sequence	
Test	**Pairs**	Digital	Paper	Paper	Digital
Linguistic Concepts	13	10.23 (2.92)	10.54 (3.04)	10.92 (1.85)	9.85 (2.12)
Recalling Sentences	17	11.35 (2.89)	12.47 (3.59)	11.12 (3.94)	11.41 (3.87)
Following Directions	18	11.00 (3.16)	11.11 (2.81)	10.33 (3.68)	10.50 (2.64)
Formulated Sentences	18	11.22 (4.14)	11.17 (3.73)	12.39 (3.60)	12.89 (3.55)

The magnitude and statistical significance of format effects are reported in Table 3. Three of the effect sizes are smaller than 0.20 in absolute value and are not statistically significant; these are within the tolerance limits for considering the formats to be equivalent. The effect size for Linguistic Concepts, however, is –0.23 *(p* < .05), indicating that performance is lower with the Q-interactive administration.

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

Test	Change Score: Digital–Paper Mean	Digital–Paper *SD*	Paper–Digital Mean	Paper–Digital *SD*	Format Effect	*t*	Effect Size
Linguistic Concepts	0.31	1.44	–1.08	2.36	–0.70	–2.05	–0.23
Recalling Sentences	1.12	1.69	0.29	1.16	–0.42	–1.78	–0.14
Following Directions	0.11	1.60	0.17	2.18	0.03	0.09	0.01
Formulated Sentences	–0.06	1.73	0.50	2.28	0.28	0.86	0.09

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

In order to investigate the cause of the format difference on Linguistic Concepts, the team compared examinees' responses shown on the video recordings with the item scores that had been automatically assigned by the Q-interactive system. The videos showed four instances for Item 9, and seven instances for Item 14, where the examinee responded correctly but the Q-interactive system assigned a score of 0. On these items the examiner asks the examinee to point to a row of pictures. Scoring was programmed so that touching one picture in the row was scored as a correct response, but touching more than one picture in the row, or touching the same picture more than once, was scored as an error. This scoring rule differs from paper-administration scoring, in which any designation by the examinee of a single row, regardless of the number of touches, is scored as correct. Instructions were provided to examiners on the capture screen and during training that they should override the system and give credit for any response that indicated the examinee was intending to point to the correct row. However, examiners evidently were not able to do that reliably.

As a result of this finding, the Q-interactive programming was changed for these items so that any number of touches of one or more pictures in a row is scored as a correct response. Assuming that this change results in correct scoring of multiple touches (which has been confirmed in beta testing), the format effect on Linguistic Concepts is eliminated. When the proper scores are assigned to Items 9 and 14 in the equivalence study data, the mean scaled scores for the digital administration change from 10.23 to 10.38 for the digital–paper sequence group and from 9.85 to 10.62 for the paper–digital sequence group. The effect size is –0.08, which is not statistically significant (*t* = –0.60).

Discussion

Two of the tests investigated in this study, Recalling Sentences and Formulated Sentences, require the examiner to transcribe accurately the examinee's oral response. For the Recalling Sentences test, examiners made edits to text shown on the examiner's tablet. For these tests, the Q-interactive

examiner interfaces worked well. The effect size of –0.14 for Recalling Sentences is comparable to that of the WIAT–III Sentence Repetition subtest (–0.08), which is a similar task using a similar examiner interface.

On the other two tests, the purpose of the study was to see how well automatic scoring of examinee touches would perform with items requiring patterns or sequences of touches. There was no difference in results on the Following Directions test between paper and Q-interactive administrations. A substantial format effect on Linguistic Concepts was traced to two items on which the automatic scoring rule diverged from the rule used in standard administration; when this difference is removed, the criterion for equivalence is met.

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.

This CELF–5 equivalence study adds to the body of evidence about the effect (or lack of effect) of features of interface design on how examiners capture and score responses, particularly for younger children. It adds new information about the accuracy of automatic scoring of examinee touches. As this body of knowledge grows, it will support generalization to other tests of the same type and features.

References

Q-interactive technical reports:

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. (2013a). *Equivalence of Q-interactive and paper administrations of cognitive tasks: Selected NEPSY®-II and CMS subtests. Q-interactive Technical Report 4.* Bloomington, MN: Pearson.

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

Daniel, M. H. (2013c). *Equivalence of Q-interactive and paper administration of WMS®–IV cognitive tasks. Q-interactive Technical Report 6.* Bloomington, MN: Pearson.

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 scales for children®–fourth edition.* Bloomington, MN: Pearson.

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

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

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

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical evaluation of language fundamentals*®–fifth edition. Bloomington, MN: Pearson.