Equivalence of Q-interactive™-Administered Cognitive Tasks: CVLT-II and Selected D-KEFS Subtests

Q-interactive Technical Report 3

Mark H. Daniel, PhD
Senior Scientist for Research Innovation

July 2012

Pearson

Introduction

Q-interactive™ is a Pearson digital platform that helps professionals give and score individually administered tests. The Q-interactive system is designed to make assessment more convenient and accurate, to give the clinician easier access to a larger number of tests, and eventually to 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 so that the examiner can 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.

The current study evaluates the equivalence of scores from digitally assisted and standard administrations of the *California Verbal Learning Test, Second Edition* (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000) and four *Delis-Kaplan Executive Function Scale* (D-KEFS; Delis, Kaplan, & Kramer, 2001) subtests: Design Fluency, Verbal Fluency, Color-Word Interference, and Trail Making.

A goal for the initial test adaptations to the Q-interactive platform was to maintain raw-score equivalence between standard (paper) and digital administration formats, so that raw scores from the two formats would be interchangeable. If equivalence is demonstrated, then the existing norms, reliability, and validity information can be applied to Q-interactive results.

This is the third equivalence study for Q-interactive. The equivalence studies of the *Wechsler Adult Intelligence Scale®, Fourth Edition* (WAIS®–IV) and the *Wechsler Intelligence Scale for Children®, Fourth Edition* (WISC®-IV) are described in Q-interactive Technical Reports 1 and 2, respectively (Daniel, 2012a, 2012b). The earlier studies found that all fifteen WAIS–IV subtests and thirteen of fifteen WISC-IV subtests yielded comparable scores in the Q-interactive and standard (paper) administrations. On two WISC–IV subtests (Matrix Reasoning and Picture Concepts), scores were slightly higher with Q-interactive administration.

In principle, digitally assisted (Q-interactive) administration may affect test scores for a number of possible reasons, including:

- examinee interaction with the tablet;
- examiner interaction with the tablet, especially during response capture and scoring; and
- global effects of the digital assessment environment.

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

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 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.

Equivalence Study Designs

Several experimental designs have been employed in Q-interactive equivalence studies. Most of them have the important feature that each examinee took a subtest only once, in either digital or standard (paper) format. This approach has the benefit of avoiding the potentially disruptive effects of taking a test twice. After an examinee has taken an item once, solving it a second time is different because they have learned the content of that item, as well as a strategy for solving that kind of problem. As a result, the cognitive processes an examinee uses during a second administration may be different from those used on the first administration. Designs that involve a single administration to each examinee avoid these problems and give examinees an experience that is similar to what they would encounter in clinical practice.

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. One of the WAIS–IV follow-up equivalence studies also used a retest design for the Processing Speed

subtests, because there was little risk of a change in cognitive process in the second administration of these tasks. The equivalent-groups and retest designs are described in detail in Q-interactive Technical Reports 1 and 2.

Another type of design, dual-capture, is appropriate for tests in which the digital format affects how the examiner captures and scores responses, but is not expected to affect examinee behavior, either directly (such as by viewing or responding on the tablet) or indirectly (by the examiner's feedback to the examinee while the examinee is performing each item). In this design, each of a relatively small number of examinees takes the test only once, but the administration is video recorded so that the examinee's responses can be seen and heard. A number of scorers independently watch each video and use either the paper or the digital format to capture and score the responses. Formats are assigned to scorings and scorers in a way that ensures that half of the scorings of each administration are in each format, and half of the scorings by each scorer are in each format, with no between-scorer correlation of formats. Mean scores using the two formats are then compared, and because the same administrations are being scored with each format, and format is balanced across administrations and scorers, the only remaining possible source of a difference between the means (other than random error) is an effect of the format on how scorers capture and score performance.

For all equivalence studies, the Q-interactive team has chosen to use an effect size of less than 0.2 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 Wechsler subtest metric that has a mean of 10 and standard deviation of 3.

Selection of Participants

The initial Q-interactive equivalence studies (including this one) 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 great 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 required amount of demographic control of 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 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. The importance in these designs is 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 and scorers participating in the equivalence studies were expected to be proficient in the test's standard administration procedures. They 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.

CVLT-II and D-KEFS Equivalence Studies

Method

The dual-capture method was used because the tests meet the requirements for this study design. On the CVLT-II and the four selected D-KEFS subtests, the examinee does not interact with the digital tablet and the examiner does not intervene while the examinee is responding to an item or trial. The examiner uses the Q-interactive interface to present instructions, time performance, and capture and score responses.

Participants

Each study (CVLT-II and D-KEFS) included 10 examinees, seven of whom participated in both studies. 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 an approximately equal number of males and females, a wide distribution of ages, ethnic diversity (at least three examinees from each of the three major groups), and diversity of socioeconomic status (education level of the examinee or the examinee's parents) with an overrepresentation of lower SES levels. Table 1 reports the characteristics of the two samples of examinees.

Table 1 Demographic characteristics of the CVLT-II and D-KEFS examinee samples

Demographic Characteristic		CVLT-II	D-KEFS
Number of Cases		10	10
Age (years)	Range	20–68	8–51
	Mean	36.6	26.6
	SD	15.8	14.8
Gender	Female	4	5
	Male	6	5
Ethnicity	African American	3	3
	Hispanic	3	3
	White	4	4
Education (or Parent Education)	< 12 years	2	2
	HS graduate	2	3
	Some post-HS	6	5
	4-year degree	0	0

Testing was conducted by examiners who were qualified and experienced in administering psychological and neuropsychological tests to children and adults and who had participated in the Q-interactive equivalence study of WAIS-IV and/or WISC-IV. The examiners received two days of onsite training in administering CVLT-II and the four D-KEFS subtests both with paper materials and with Q-interactive, and they conducted several practice administrations before the study began.

After completing their training and practice administrations and doing one or several administrations for use in the study, examiners served as scorers. Examiners/scorers who were not Pearson employees were compensated for their participation.

Measures

CVLT-II

The CVLT-II is used to assess recall and recognition of items in a list of 16 nouns that belong to four categories (furniture, vehicles, vegetables, and animals). It includes the following types of free-recall and cued-recall scores:

- immediate free recall: the list is presented five times in a standard sequence, and after each presentation the examinee is asked to say as many of the words as possible; the total score is the sum of correct responses on the five presentations

- short-delay free and cued recall: after one administration of an interference task (free recall of a different list of 16 words), the examinee is asked to recall words from the initial list, first with no cuing, and then after being told each of the four categories

- long-delay free and cued recall: about 20 minutes later, and without forewarning, the examinee is asked to recall words from the initial list, both without and with category cuing

During free-recall and cued-recall tasks, the examiner records each response. With Q-interactive, this is done either by handwriting (using fingertip or a stylus) or by touching a “picklist” button. An audio recording is made of the responses, which the examiner may listen to post-administration to review responses and clarify scoring. Either during or after the administration, the examiner classifies each response as correct, a repetition, or an intrusion (i.e., a word that was not on the list).

In addition, two recognition trials (yes/no and forced-choice) are administered after the long-delay cued recall trials. The examiner records responses simply by marking or touching the word that the examinee said.

In this study, all portions of the CVLT-II were administered, but only the free-recall trials were evaluated. The cued-recall trials have the same capture interface as the free-recall trials, and the recognition trial responses are extremely easy to record and score and do not raise format-equivalence questions.

D-KEFS Subtests

D-KEFS Design Fluency: The examinee has a sheet of paper with rows of squares. Within each square are small circles in a random pattern; in Condition 1 all of the circles are black, and in Conditions 2 and 3 half of the circles are solid black, and half are white with a black outline. On the first one-minute condition, the examinee creates as many distinct designs as possible by connecting solid circles with four continuous straight lines. On the second condition, the task is

the same but the examinee connects the white circles. On the third condition, the examinee must alternate between black and white circles. Scoring is done post-administration and based on the number of unique designs on each trial.

D-KEFS Verbal Fluency: On each of six one-minute trials, the examinee is asked to say as many distinct words as possible that meet a certain criterion. For the first three trials, the words must begin with a particular letter, for the next two trials, the words must belong to a particular semantic category, and for the last trial, words must alternate between two semantic categories. During each trial, the examiner writes the responses in one of four fifteen-second response areas. With Q-interactive, there is a single handwriting-capture area and the word is automatically transferred to the appropriate response area according to the elapsed time. With Q-interactive, the examiner also has the option of listening to an audio recording of the responses post-administration and writing them in the appropriate fifteen-second areas. The score is the number of distinct responses that conform to the rules (e.g., no names of people or places), and scoring is done post-administration.

D-KEFS Color-Word Interference: On each condition, the examinee names, as quickly as possible, visually presented stimuli presented in rows. On the first condition, the stimuli are color patches (red, green, and blue). On the second condition, the stimuli are the words “red,” “green,” and “blue.” On the third condition, the stimuli are the color words printed in ink colors that do not match the word, and the examinee names the ink colors. On the final condition, the stimuli are the same as on the third condition except that some words are enclosed in a box; the examinee names the ink color if the word is not in a box, and says the word if it is in a box. For both standard and Q-interactive administrations, the stimuli are printed in a stimulus book. (The stimulus page is too large to be displayed on the tablet screen.) During administration, the examiner marks errors and records the completion time. The primary score on each condition is based on completion time, and secondary scores are based on errors.

D-KEFS Trail Making: The examinee responds by drawing in a printed booklet. On the first condition, the examinee crosses out dots that have a particular number inside them. On the second condition, the examinee draws lines connecting numbered dots, in numerical sequence. On the third condition, the examinee does the same thing, but the dots are labeled with letters rather than numbers. On the fourth condition, half of the dots are numbered and half have letters, and the examinee connects them by alternating between numbers and letters. The fifth condition is a simple motor-speed task in which the examinee traces lines connecting dots. During administration, the examiner tells the examinee to correct any errors as they occur, and records the completion time. The primary score on each condition is based on completion time and secondary scores are based on errors. An additional score is a composite of scores on the number-sequence and letter-sequence trials.

Procedure

Training, testing, and scoring took place at Pearson’s office in San Antonio, TX in May and June 2012. Each administration used in the studies was conducted using the standard (paper) materials. For examinees participating in both studies, the D-KEFS Trail Making and Design Fluency subtests were administered during the CVLT-II long-delay period, and D-KEFS Verbal Fluency and Color-Word Interference were administered after the CVLT-II recognition trials.

In each study (CVLT-II and D-KEFS), the ten administrations were conducted by three examiners who each did three or four administrations. In order to make the administration as realistic as possible, the examiner performed all of the usual recording and scoring steps; 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 response booklet (if any), but not showing the examiner's record form.

Standard administration procedures were followed with two exceptions. First, for the D-KEFS Verbal Fluency subtest, the examiner did not stop a trial until about ten seconds after the 60-second mark. This was done to enable the scorers to apply their own timing. Second, examinees were asked to make certain kinds of errors on several of the subtests or trials, so that these errors would be represented in the study. The errors in question were expected to be relatively challenging for scorers to handle, but do not occur frequently in practice. On the Switching condition of D-KEFS Trail Making, several examinees were asked to make set-loss or sequencing errors; and on CVLT-II, several examinees were asked to give intrusions or repetitions on the free-recall trials.

Each administration was scored by ten scorers; five used the Q-interactive format and five used the paper format. Only these post-administration scorings were used in the analyses. In each study, each scorer scored all ten administrations, five in each format. To ensure that there was no correlation of formats between scorers, format assignments were made using a random number table to identify the five 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 on a monitor, and recorded responses on either the paper record form or the Q-interactive tablet. For D-KEFS subtests that use a response booklet, the video used picture-in-picture with a small window showing the examinee superimposed on the full screen showing the response booklet from the examiner's point of view. On those subtests, the scorer had a copy of the completed response booklet to use for scoring. 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, they enabled the researchers to investigate possible causes by reviewing the scorers' behaviors; and second, they provided information about how scorers interact with the digital and paper materials, which can be helpful in future test design. These videos were shot from behind the scorer and showed the monitor that the scorer was watching and the tablet or record form on which the scorer was capturing responses.

Pearson staff reviewed the completed record forms from paper scorings for score-calculation accuracy and corrected any errors in combining item or trial scores. There were no such errors in the Q-interactive scorings. The raw information recorded by the scorers during scoring was kept intact and was not modified during this review.

The method of analysis was to compare paper and Q-interactive mean normative scores (scaled scores, *T* scores, or *z* scores) for each score variable of interest. Because the ten administrations in each study were each scored ten times, five times with each format, there would be 50 paper-format scores and 50 Q-interactive scores in each comparison. The effect size for each score variable is the difference in means divided by the population standard deviation of the score's

normative metric (e.g., 3 for a scaled score). A *z* test was applied to evaluate statistical significance, using the root mean square of the deviations of scores from the within-administration-and-format means as the standard error.

Several of the tests and subtests in these studies produce a large number of scores, of which only the primary scores were analyzed. In addition, scores for which the response-capture process does not place any plausible demands on the Q-interactive interface were not analyzed (e.g., CVLT-II Long Delay Recognition, in which the responses are simply “Yes” and “No”).

Results

The goal of having five scorings in each format was achieved for most of the tests and subtests, but for a few, there were only four scorings in one or both of the formats. In order to avoid the bias that would result from having an administration represented an unequal number of times in the overall mean scores for the two scoring formats, weighted means were used in which the mean score (across scorings) for each of the ten administrations was given an equal weight. Table 2 reports these weighted means and standard deviations of CVLT-II and D-KEFS scores in each scoring format, and the effect size of the difference between them.

Table 2 Differences between scores obtained using paper and Q-interactive recording formats

Test/subtest and score	Paper scoring			Q-interactive scoring			Difference	*z*	Effect size
	N	Mean	*SD*	*N*	Mean	*SD*			
CVLT-II									
Immediate Free Recall: Total (*T*)	47	54.99	11.52	48	55.06	11.01	0.07	0.07	0.01
Short-delay Free Recall: Total (z)	48	0.31	0.99	48	0.26	0.99	-0.05	-0.53	-0.05
Long-delay Free Recall: Total (z)	48	0.16	0.94	48	0.20	0.88	0.04	0.37	0.04
D-KEFS (scaled scores: *M*=10, *SD*=3)									
Trail Making									
Visual Scanning	50	11.04	1.96	50	11.00	2.05	-0.04	-0.06	-0.01
Number Sequencing	50	12.42	1.63	50	12.36	1.66	-0.06	-0.19	-0.02
Letter Sequencing	50	11.42	2.52	50	11.36	2.60	-0.06	-0.17	-0.02
Number-Letter Switching	50	10.00	2.60	50	9.96	2.59	-0.04	-0.32	-0.01
Motor Speed	50	11.82	1.10	50	11.76	1.22	-0.06	-0.14	-0.02
Combined Letter & Number Seq.	50	12.70	2.12	50	12.60	2.18	-0.10	-0.26	-0.03
Verbal Fluency									
Letters Correct	49	11.24	2.68	49	11.12	2.52	-0.12	-0.26	-0.04
Categories Correct	50	12.06	3.63	48	11.81	3.66	-0.25	-0.52	-0.08
Design Fluency									
Filled Dots	50	10.27	2.65	50	10.36	2.47	0.09	0.22	0.03
Empty Dots	50	10.06	3.14	50	10.04	3.43	-0.02	-0.03	-0.01
Switching	50	10.27	2.55	50	10.24	2.64	-0.03	-0.04	-0.01
Composite	50	10.70	3.10	50	10.66	2.98	-0.04	-0.07	-0.01
Color-Word Interference									
Color Naming	50	10.66	2.89	47	10.49	2.98	-0.17	-0.45	-0.06
Word Reading	50	9.86	3.31	48	9.60	3.04	-0.26	-0.63	-0.09
Inhibition	50	10.78	2.81	48	10.79	2.81	0.01	0.03	0.00
Inhibition/Switching	50	10.26	3.06	48	10.09	3.04	-0.17	-0.54	-0.06

Mean scores were slightly higher than the population averages, and standard deviations were close to the population averages on all tests except D-KEFS Trail Making where they were relatively small. With this exception, the examinee samples provided representative distributions of performance levels. All effect sizes were very small and non-significant: The median absolute value was 0.02, and the greatest was 0.09.

Discussion

These are the first Q-interactive equivalence studies to focus exclusively on the possible effect of using the digital interface to capture and score responses. The CVLT-II and the four D-KEFS subtests that were analyzed are tests on which the only plausible sources of format effect would be in the recording and scoring process, because the examinee does not interact with a tablet and the examiner does not intervene while the examinee is responding. On CVLT-II free-recall trials and D-KEFS Verbal Fluency and Color-Word Interference, recording responses accurately is challenging regardless of the format, because examinees often respond at a rapid pace. These score variables were of particular interest in these studies, and it is valuable to find no difference in the accuracy of scoring between paper and Q-interactive methods.

The design of these studies, with multiple scorings of a small number of administrations, did not permit evaluation of whether format effects might be different as a function of examinee characteristics (age, gender, ethnicity, or socioeconomic status). The previous Q-interactive equivalence studies found a slight tendency toward a positive digital-format effect for younger examinees. However, because examinees do not interact with the digital materials, such variation with age would not be expected on CVLT-II or these D-KEFS subtests.

The CVLT-II and initial D-KEFS equivalence studies add to the body of evidence about the effects (or lack of effect) of features of interface design on how examiners capture and score responses. As this body of knowledge grows, it should support generalization to other tests of the same type and features.

References

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.

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.

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.