Equivalence of Q-interactive™ and Paper Administrations of Cognitive Tasks: WISC®–V

Q-interactive Technical Report 8

Mark H. Daniel, PhD

Dustin Wahlstrom, PhD

Ou Zhang, PhD

September 2014

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 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 norms, reliability, and validity information gathered for the paper format can be applied to Q-interactive results.

This is the eighth Q-interactive equivalence study. In this study, the equivalence of scores from digitally assisted and standard administrations of the *Wechsler Intelligence Scale for Children*®–fifth edition (WISC®–V; Wechsler, 2014) 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 (i.e., 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 next study, all subtests of the *Wechsler Memory Scale*–fourth edition (WMS–IV; Wechsler, 2009b) were found to be equivalent. Finally, the Recalling Sentences, Formulated Sentences, Following Directions, and Linguistic Concepts tests that are part of the *Clinical Evaluation of Language Fundamentals*–fifth edition (CELF–5; Semel, Wiig, & Secord, 2013) were evaluated. The first two tests, which require the examiner to record and score an extended oral response, were found to be equivalent. The last two subtests were the first to implement automatic scoring of the examinee's touch response. This yielded equivalent scores on Following Directions, but not on Linguistic Concepts, where certain patterns of examinee touches on two items were improperly scored as errors. When this programming error was corrected, equivalence was demonstrated.

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) have up to now been used with the Q-interactive administration. The eventual goal is to replace such components by interactive digital interfaces, but this will involve switching from a strategy of demonstrating equivalence (which provides a basis for relying on existing psychometric evidence) to one of demonstrating the validity and reliability of the digital version. This is being undertaken for the first time with the WISC–V processing speed subtests, but that work is still in progress and is not described in this report, which focuses solely on equivalence.

- 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.

- How accurately the Q-interactive system captures and scores the examinee's touch responses. Q-interactive first introduced automatic scoring with the CELF-5. Previous implementations 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 assign scores to touch 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, or even give shorter responses, so as to avoid having to wait for examiners who were slower typists to finish keying in their verbatim responses. Another type of global effect was observed with some very young children (aged 2 or 3) who became distracted by the tablet which they perceived as a toy; this problem is being addressed through research into the human factors issues at these ages.

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. A reasonable objective for a new technology is for it 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.

In the previous equivalence studies most or all administrations were video recorded, showing the examiner's and examinee's interactions with their tablets. This provides a way to check the accuracy of administration, recording, and scoring in both digital and standard formats if a format effect (i.e., non-equivalence) is found, so that the cause of the difference can be identified and corrected. These recordings had the additional benefit of showing how examiners and examinees interacted with the test materials in each format, information that can be used to guide improvements in interface design. Given the extensive experience that the Q-interactive team has acquired over the first seven studies spanning two years and over a thousand administrations, the WISC–V administrations in this study were not video recorded.

As a whole, the equivalence studies indicate that examinees ages 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, on the WISC–IV Matrix Reasoning and Picture Concepts subtests, suggested that on subtests involving conceptual reasoning with detailed visual stimuli, children may perform better when the stimuli are shown on the tablet. (The current study of the WISC–V provides an opportunity to examine the replicability of this finding.) Also, the cumulative evidence indicates that when examiners use the various Q-interactive interfaces, they obtain the same results as with the paper materials.

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 the possibility that the way in which an examinee interacts with the task will change 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 each examinee takes the test only once closely approximate 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. This design, with random assignment, has been used for the study of the WISC–V.

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 powerful 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), WMS–IV (Technical Report 6), and CELF–5 (Technical Report 7).

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 compromise the interpretability of 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, race/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.

WISC–V Equivalence Study

Method

Measures

The WISC–V is a comprehensive instrument used to assess intellectual ability at ages 6 to 16. Many of its subtests are nearly identical in administration and scoring to ones in the WISC–IV, which has already been evaluated for equivalence of Q-interactive and paper formats. However, because several WISC–V subtests are new, and some others have had minor changes in administration or scoring (including changes to the Q-interactive examiner interface), the entire battery has been re-evaluated for equivalence.

Digital versions of the three Processing Speed subtests (Coding, Symbol Search, and Cancellation), in which the examinee responds by touching or drawing on the tablet, were administered in the WISC–V equivalence study. In order to minimize construct-irrelevant differences between the paper and digital versions, the paper materials were reformatted from their WISC-IV versions to resemble the Q-interactive format. Nevertheless, it was assumed that the paper and digital versions would not be raw-score equivalent because of the difference in response mode. The goal of the development project was for the two versions of each subtest to measure the same construct and to have similar psychometric properties so that they could be equated and could be treated as alternate forms for purposes of clinical interpretation. The data indicated, however, that the correspondence between the paper and digital versions was not yet close enough to support these goals. Further design and development work on these subtests is underway. The initial release of the WISC–V continues to use paper response booklets for the Processing Speed subtests, with a Q-interactive examiner interface for timing and recording that is the same as the one that has been shown to be equivalent to the paper format in the studies of the WISC-IV and the WAIS–IV. Because the examinee works in a response booklet and the simple examiner interface is identical to that of WISC-IV, these subtests were not re-evaluated.

Participants

This study was carried out as part of the WISC–V standardization. Almost all administrations were conducted in April and May of 2014, with a small number of Q-interactive administrations conducted in the preceding several months. The examinee sample consisted of nonclinical children aged 6 to 16. Pearson's Field Research staff recruited examinees and compensated them for their participation. Potential participants were screened for demographic characteristics and exclusionary factors, such as perceptual or motor disabilities or clinical conditions. The sampling plan called for approximately equal numbers of males and females and approximately equal numbers at each year of age, and representative proportions of racial/ethnic groups and levels of socioeconomic status (parent education level).

Potential participants whose characteristics matched the demographic requirements of both the paper and Q-interactive samples were randomly assigned to one of the formats. After data collection was completed, the research team selected pairs of cases (one member from each format) matched by age range, gender, ethnicity, and parent education to use in the analysis.

Examiners were based at numerous locations around the country. All examiners received remote training on the WISC–V and on Q-interactive administration so that they were eligible to administer the test in either format. They conducted practice administrations and received feedback on any administration errors. Examiners who were not Pearson employees were compensated for their participation.

Procedure

Examiners captured response information in the standard manner used for norming, which includes writing the complete verbatim response to each verbal subtest item, and scored all items. The Pearson research team checked paper-administration cases for proper use of administration rules (such as querying, start points, and discontinue rules), but did not rescore items. All subtest raw scores were calculated automatically, either by Pearson staff using the keyed item scores, or by the Q-interactive system.

The format effect on each subtest was estimated using a multiple regression approach in which the independent variables were demographics (age, gender, ethnicity, and parent education), scores on selected other WISC–V subtests or indexes, and format, and the dependent variable was the subtest scaled score. Although the use of demographically-matched examinee pairs with random assignment to format provides good experimental control, the use of covariates further enhances statistical power. WISC–V subtests rather than external tests were used as covariates because the WISC–IV study had already shown that, with only a few exceptions, the subtests are minimally affected by format. The analysis was carried out in two stages. The first stage used as predictors the following subtests or indexes which had shown small format effects in the WISC–IV equivalence study: Block Design, Arithmetic, Digit Span, and the Verbal Comprehension Index. (For analysis of each verbal subtest, the VCI was replaced by the Similarities or Vocabulary subtest.) The results of this stage were used to refine the set of predictors, and the analysis was repeated in stage two.

Because format was coded 0 for paper and 1 for Q-interactive, the unstandardized regression weight for format was a direct estimate of the format effect in the normative-score units for that subtest. This was converted to an effect size by dividing by the normative-score standard deviation (either 3 or 15). Assuming a multiple correlation of .5 between the set of demographic and subtest-score predictors and the score on the target subtest, a total sample of 350 cases has power of .56 to detect an effect size of 0.20 (alpha = .05).

In order to determine whether there might be a format effect for subgroups of the population by age, gender, SES, ethnicity, or ability level, a format effect indicator was calculated for each Q-interactive examinee and the relationship of this indicator to each of the individual characteristics was analyzed. The format effect indicator was the difference between the examinee's actual subtest score (with a digital administration) and the paper-format score predicted for that examinee on the basis of the demographic variables and other WISC–V subtest scores.

Results

A total of 350 examinees (175 matched pairs) were selected for analysis. Of the 113 participating examiners, 58% submitted cases in both formats, 38% submitted only Q-interactive administrations, and 4% submitted only paper administrations. Table 1 reports the characteristics of the examinee

samples taking the paper and Q-interactive administrations. The two samples resemble each other closely on each demographic variable. Overall, the sample has a nearly equal representation of each age, and has nationally-representative proportions of ethnic groups. Females, and children of college graduates, are somewhat overrepresented relative to the general population.

Table 1 Demographic characteristics of the samples

		Administration Format	
Demographic Characteristic		**Paper**	**Q-interactive**
Number of Cases		175	175
Age (years)	6	14	16
	7	18	15
	8	15	16
	9	15	15
	10	12	12
	11	20	20
	12	14	14
	13	17	17
	14	18	18
	15	16	16
	16	16	16
	Mean	11.1	11.1
	SD	3.2	3.2
Gender	Female	101	102
	Male	74	73
Ethnicity	African American	18	19
	Hispanic	29	32
	White	117	115
	Other	11	9
Parent Education	<12 years	17	12
	HS graduate	30	29
	Some post-HS	54	55
	4-year degree	74	79

Table 2 reports the means and standard deviations of scores on the WISC–V subtests for each format. Given the close similarity of the demographic characteristics of the two format groups and the fact that examinees were randomly assigned to format, one would not expect large or systematic differences in scores between the groups.

Table 2 Descriptive statistics for WISC–V subtests, by administration format

Subtest	Paper		Q-interactive	
	Mean	*SD*	Mean	*SD*
Arithmetic	10.3	2.6	9.9	2.4
Block Design	9.9	2.5	10.6	2.5
Comprehension	9.9	2.6	9.5	2.6
Digit Span	10.1	2.7	10.6	2.5
Figure Weights	10.0	2.6	10.6	2.9
Information	10.2	3.0	10.2	2.6
Letter-Number Sequencing	10.3	2.5	10.7	2.4
Matrix Reasoning	9.9	2.5	10.6	2.9
Picture Concepts	9.9	2.9	10.1	3.2
Picture Span	10.3	2.5	10.7	2.7
Similarities	10.2	2.8	10.3	2.6
Visual Puzzles	9.8	2.6	10.0	2.7
Vocabulary	10.0	3.0	9.8	2.6
Immed. Symbol Translation	99.4	13.4	100.7	13.2
Delayed Symbol Translation	99.9	13.2	100.8	13.9
Recog. Symbol Translation	101.7	12.6	102.5	13.2
Naming Speed Literacy	100.7	13.9	103.0	14.2
Naming Speed Quantity	101.4	14.5	102.2	12.6

Results from the first stage of the regression analysis led to selection of a new set of predictor variables with relatively low format effects for use in the second stage: Visual Puzzles, Picture Concepts, Picture Span, Letter-Number Sequencing, and the Verbal Comprehension Index (just Similarities or Vocabulary for analysis of the verbal subtests). Table 3 shows the results of the second set of regression analyses, including the multiple correlation with the predictor variables, the unstandardized regression weight for format, the *t* value and statistical significance associated with format as a predictor, and the effect size for format. Results from the WISC–IV equivalence study are shown for comparison.

Multiple correlations ranged from .35 to .71 (median = .52), indicating that the demographic and ability variables accounted for, on average, about one-fourth of the variance in the subtest score. The purpose of the predictor variables was to increase the power of the analyses by reducing the amount of variance to be explained; thus, high multiple correlations are desirable, but are not essential for the validity of the analyses.

Table 3 Effect size of Q-interactive format on each WISC–V subtest

Subtest	*R*	Unstand. Regression Weight	*t*	Effect Size	WISC–IV Effect Size
Arithmetic	.55	–0.49	–2.11*	–0.16	0.10
Block Design	.58	0.59	2.66**	0.20	0.02
Comprehension	.55	–0.59	–2.51*	–0.20	0.00
Digit Span	.54	0.25	1.04	0.08	0.13
Figure Weights	.52	0.49	1.95	0.16	—
Information	.71	–0.15	–0.68	–0.05	0.07
Letter-Number Sequencing	.50	0.26	1.13	0.09	0.18
Matrix Reasoning	.48	0.51	1.99	0.17	0.27
Picture Concepts	.40	0.07	0.22	0.02	0.21
Picture Span	.42	0.21	0.83	0.07	—
Similarities	.66	0.11	0.50	0.04	0.02
Visual Puzzles	.52	0.11	0.46	0.04	—
Vocabulary	.66	–0.39	–1.69	–0.13	0.05
Immed. Symbol Translation	.57	0.52	0.44	0.03	—
Delayed Symbol Translation	.50	0.21	0.16	0.01	—
Recog. Symbol Translation	.44	–0.07	–0.06	0.00	—
Naming Speed Literacy	.41	1.73	1.24	0.12	—
Naming Speed Quantity	.35	–0.37	–0.27	–0.02	—

Note: A positive effect size indicates higher scores with Q-interactive. The unstandardized regression weight is in the subtest's normative score metric (*SD* of 3 for the first 13 subtests, and 15 for the last five).
*$p < .05$, **$p < .01$

Three of the eighteen subtests showed a statistically significant format effect at the .05 level, but none of the effect sizes exceeded the criterion of 0.20 that is the standard in the Q-interactive equivalence studies. Among the ten subtests common to WISC–IV and WISC–V, the pattern of effect sizes was moderately similar in the two studies ($r = .44$), but there were several substantial differences. The two subtests with the largest effect sizes on WISC–V, Block Design and Comprehension, had had effect sizes near 0 in the WISC–IV study. On the other hand, the two subtests that showed effect sizes greater than 0.20 in the WISC–IV study had moderate (Matrix Reasoning, 0.17) to small (Picture Concepts, 0.02) effect sizes in the WISC–V study. Furthermore, whereas all of the WISC–IV effect sizes (excluding processing speed) were positive, the effect sizes in the WISC–V study were more balanced (five negative and thirteen positive).

The possible existence of differential format effects for examinees of different ability levels or demographic characteristics was investigated in a series of analyses. A format effect indicator was calculated for each examinee who was administered a subtest via Q-interactive. This indicator was the difference between the examinee's actual subtest score and their predicted paper-administration score, using the set of demographic variables and subtest scores from the stage two analyses as predictors. For the continuous predictor variables (ability, age, and SES) the measure of relationship was the correlation coefficient between each variable and the format effect indicator; for gender, it was the *t* statistic; and for ethnicity, it was the *F* statistic from analysis of variance. In this analysis, the ability score was the predicted paper-administration score. Results are shown in Table 4.

Table 4 Relationship of format effect to ability level and demographics

	Correlation			Gender	Ethnicity
Subtest	**Ability[a]**	**Age**	**SES**	***(t)*[b]**	***(F)***
Arithmetic	–.01	.03	–.05	–0.40	0.06
Block Design	.02	.04	.03	0.08	0.67
Comprehension	.00	–.05	–.02	0.63	0.88
Digit Span	–.03	.06	–.01	–0.80	0.33
Figure Weights	.03	–.03	–.04	–0.46	0.42
Information	–.10	–.07	–.04	–0.21	0.78
Letter-Number Sequencing	–.14	.03	–.06	0.83	1.08
Matrix Reasoning	.02	–.01	–.01	0.55	0.17
Picture Concepts	.04	.04	–.02	–0.06	0.59
Picture Span	.04	.05	.09	–0.20	0.66
Similarities	.04	.07	–.04	–0.60	0.47
Visual Puzzles	.09	–.07	.11	0.24	0.91
Vocabulary	–.11	–.01	–.07	0.37	0.69
Immed. Symbol Trans.	–.02	–.06	–.01	–0.70	1.26
Delayed Symbol Trans.	–.02	–.10	–.07	–0.35	2.17
Recog. Symbol Trans.	–.02	–.07	–.01	–0.13	0.98
Naming Speed Literacy	–.10	.03	–.08	–1.80	0.10
Naming Speed Quantity	–.06	.15*	–.12	–2.00*	0.64

Note. *N* = 175. For this analysis, the format effect is the difference between the actual Q-interactive subtest score and the score predicted for a paper administration from demographics and other subtests.

[a] Predicted paper-administration subtest score.

[b] A positive value of *t* means that the format effect was greater for females.

*$p < .05$.

Among 90 statistical tests one would expect to obtain four or five statistically significant results (at the .05 level) simply by chance, yet only two of the results shown in Table 4 were statistically significant. Both were for the Naming Speed Quantity subtest, where there was a tendency for the Q-interactive administration to benefit older rather than younger examinees, and males rather than females. Overall, the findings reported in Table 4 reinforce the comparable finding in the WISC-IV study that there are no differential effects of the Q-interactive format by ability level, age, socioeconomic status, gender, or ethnicity.

Discussion

This study was a replication of the WISC-IV equivalence study for ten of the WISC–V subtests, as well as the initial study of eight new subtests. Consistent with the overwhelming pattern of results in the seven previous equivalence studies, all format effect sizes fell within the established criterion for Q-interactive equivalence (i.e., 0.20 or less). Two subtests that had shown effect sizes greater than 0.20 on the WISC–IV (Matrix Reasoning and Picture Concepts) had smaller effect sizes in this study. On the two subtests that showed the largest effect sizes in the WISC–V study (+0.20 for Block Design and –0.20 for Comprehension), there is minimal or no interaction of the examinee with the tablet: on Block Design the examinee tablet simply displays the design to be created, and on Comprehension the examinee tablet is not used at all. Furthermore, these subtests had effect sizes of only 0.02 and 0.00, respectively, in the WISC–IV study.

Over the course of the series of Q-interactive equivalence studies it becomes clear that although small effects may be observed in some studies, unless a cause can be detected (such as by studying the video recorded administrations), these effects generally are not replicable or systematic and are likely due largely to the fact that there is no perfect research design for evaluating these issues without error. Nevertheless, on several occasions the studies identified genuine flaws that were corrected before publication but which might otherwise have gone undetected (e.g., the error in the original timing function for the WAIS–IV processing speed subtests, and the subtle image-quality problem on WAIS–IV Picture Completion). It is the combination of empirical results and close scrutiny of administration and scoring procedures that provides a level of confidence in equivalence that could not be obtained from either method alone.

The current study also showed that there were virtually no statistically significant differences in format effect among subgroups by age, gender, ethnicity, socioeconomic status, or ability level. This has been the consistent finding of previous Q-interactive equivalence studies as well. This set of results indicates that the general finding of an absence of format effects applies broadly to the general nonclinical population.

References

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

Daniel, M. H. (2012a). *Equivalence of Q-interactive administered cognitive tasks: WAIS®–IV.* (Q-interactive Technical Report 1). Bloomington, MN: Pearson.

Daniel, M. H. (2012b). *Equivalence of Q-interactive administered cognitive tasks: WISC®–IV* (Q-interactive Technical Report 2). Bloomington, MN: Pearson.

Daniel, M. H. (2012c). *Equivalence of Q-interactive administered cognitive tasks: CVLT®–II and selected D-KEFS® subtests* (Q-interactive Technical Report 3). Bloomington, MN: Pearson.

Daniel, M. H. (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.

Daniel, M. H., Wahlstrom, D., & Zhou, X. (2014). *Equivalence of Q-interactive® and paper administrations of language tasks: Selected CELF®–5 tests* (Q-interactive Technical Report 7). 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.

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

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

Wechsler, D. (2014). *Wechsler intelligence scale for children®*–fifth edition. Bloomington, MN: Pearson.