Equivalence of Q-interactive™-Administered Cognitive Tasks: WISC®–IV

Q-interactive Technical Report 2

Mark H. Daniel, PhD

Senior Scientist for Research Innovation

July 2012

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 *Wechsler Intelligence Scale for Children®, Fourth Edition* (WISC®–IV; Wechsler, 2003). 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 second test instrument to be adapted to Q-interactive. The equivalence studies of the first instrument, the *Wechsler Adult Intelligence Scale®, Fourth Edition* (WAIS®–IV; Wechsler, 2008), are described in Q-interactive Technical Report 1 (Daniel, 2012). That research found that all fifteen WAIS–IV subtests yielded comparable scores in the Q-interactive and standard (paper) administrations. In view of the close similarity of subtest content and formats between WAIS–IV and WISC–IV, it was anticipated that similar results would be obtained in the study of WISC–IV.

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, the physical manipulatives (Block Design blocks) and printed response booklets (Processing Speed subtests) of the WISC–IV and WAIS–IV were used with the Q-interactive administration. Though these physical components may eventually be replaced by interactive digital interfaces, the degree of adaptation required would make raw-score equivalence unlikely, which means that more extensive development efforts would be required 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 that are 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 is 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 used a (randomly or non-randomly) equivalent-groups design, in which each examinee took a subtest only once, in digital or standard (paper) format. This design requires larger samples than a retest or alternate-form design, but it has the important benefit of avoiding the potentially disruptive effects of taking a test twice. Some of the WISC–IV subtests have practice effects when administered a second time in the standard format. More importantly, 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. So, the cognitive processes an examinee uses during a second administration may be different. An equivalent-groups design that compares the performance of two groups, one taking the test in the digital format and the other in the paper format, avoids these problems and gives examinees an experience that is highly similar to what they would encounter in clinical practice.

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.

Randomly Equivalent Groups Design

Randomly equivalent groups is the design used for the WISC–IV equivalence study. With this design, the sample should resemble the general population in terms of sex, ethnicity, and education level. The distribution of ages should reflect the research questions (e.g., over-representing age levels at which a particular risk of nonequivalence is suspected). Within each demographic cell (combination of sex, ethnicity, and education, by age), half of the examinees are randomly assigned

to a test format. Examiners also are randomly assigned to different formats for different examinees. Immediately following test administration, all examinees take a set of covariate tests in paper format that measure the same construct(s) as the digitally administered test (the focal test).

The results of each focal test or subtest are then analyzed separately, using multiple regression or ANCOVA. In the regression method, the predictors are age-adjusted normative scores on the covariate tests, demographic variables, and a dummy-coded variable that represents administration format. The dependent variable is the age-adjusted normative score on the focal test. The unstandardized regression weight for format is a measure of the format effect, expressed in the focal test's normative-score metric. Dividing the average format effect by the standard deviation of the normative-score metric yields the effect size.

The advantage of the randomly equivalent groups design is that the random assignment of examinees to format tends to make the subsamples being compared equivalent on all characteristics that may influence test performance, including those that are not measured (or cannot be measured). This advantage comes at the price, however, of requiring a relatively large sample. For example, if the combination of demographics and the covariate tests has a multiple correlation of 0.7 with the score on the test being analyzed, then obtaining power of 0.8 to detect an effect size of 0.2 (with alpha of .05) would require 200 examinees per format.

Non-Randomly Equivalent Groups Design

In this design, the existing norm sample serves as the paper-administration sample and only the digital-administration sample must be collected. This method leverages the large and carefully stratified norm sample that exists for each test. It can be used when the focal test contains two or more subtests that measure the same ability construct (so that they can serve as covariates for one another), or when the norm sample examinees took external covariate tests that can also be administered to the Q-interactive sample.

This design reduces the number of cases that need to be collected, but it foregoes the benefits of random assignment of examinees to format. It was used for the WAIS–IV equivalence studies, as described in detail in Q-interactive Technical Report 1.

Other Designs

Occasionally, the nature of a test lends itself to a more efficient type of design in which examinees serve as their own controls, such as retest and dual-capture. (The alternate-form design has not been feasible because the WISC–IV and WAIS–IV subtests do not have alternate forms.) Each of these designs is described more fully in Q-interactive Technical Report 1.

Retest Design

In the retest design, each examinee takes the test twice, once in each format, and the administration sequence is counterbalanced; half the examinees take one format first and half take the other first. This design is appropriate when the response processes are unlikely to change substantially on retest, because the examinee does not learn solutions to specific problems or strategies for solving novel problems. Examples of such tests are measures of processing speed, or of short-term memory for non-meaningful stimuli.

Dual Capture Design

In the dual-capture design, each examinee takes the test only once, but the administration is video recorded to capture the examinee's responses and all audio. A number of examiners independently watch each video to capture and score the responses, using either the paper or the digital format according to a random assignment. This design is appropriate for subtests where the digital format does not 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). The design focuses entirely on the effect of the digital format on the examiner's accuracy in capturing and scoring performance.

Selection of Participants

The initial Q-interactive equivalence studies used samples of nonclinical examinees with demographic characteristics similar to those of the general population. Examinees with clinical conditions were excluded in order to focus the studies on estimating the presence and size of any format effects. Because the effects of digitally assisted administration on individuals with particular clinical conditions are difficult to predict, including an arbitrarily determined sample of examinees with various disorders would have unknown effects on the results and could interfere with the goal of seeing whether the digital format has an effect on examinee or examiner behavior.

Examiners 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 procedures to be able to conduct the administration 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–IV Equivalence Study

Method

The randomly equivalent groups method was used for the WISC–IV equivalence study. Data was collected between December 2011 and March 2012.

Participants

The sample consisted of 344 children, ages 6 to 16, who were recruited by Pearson's Field Research staff or by a market research company. Potential examinees were screened for demographic characteristics and exclusionary factors, such as perceptual or motor disabilities or severe clinical conditions. The sampling plan required an even distribution of cases across ages, though with more cases per year at the younger ages and fewer cases at the older ages. Gender, ethnicity, and socioeconomic status (parent education level) reflected the national population within each year of age. All examinees were paid for their participation.

Table 1 reports the demographic characteristics of the sample. The subgroups taking WISC–IV with the standard or Q-interactive format were very similar. Overall, there was nearly equal representation of males and females. Hispanic children were overrepresented (31% as opposed to 25% in the child population) and white children were under-represented (49% rather than 56%). The percentage of children whose parents had some education beyond high school (72%) was greater than in the general population (64%).

Table 1 Demographic characteristics of the WISC–IV sample

Demographic Characteristic		Administration Format	
		Standard (paper)	Q-interactive
Number of Cases		175	169
Age (years)	6	15	15
	7	15	17
	8	18	15
	9	23	17
	10	23	24
	11	17	19
	12	15	14
	13	11	18
	14	17	9
	15	11	11
	16	10	10
	Mean	10.6	10.6
	SD	3.0	2.9
Gender	Female	53%	51%
	Male	47%	49%
Ethnicity	African American	12%	13%
	Asian	3%	4%
	Hispanic	29%	33%
	White	51%	46%
	Other	5%	5%
Parent Education	< 9 years	2%	4%
	9–11 years	8%	7%
	HS graduate	19%	17%
	Some post-HS	34%	32%
	4-year degree	37%	40%

Examiners were school and clinical psychologists qualified and experienced in administering WISC–IV. Most of them had participated in the WAIS–IV equivalence study and had used Q-interactive. All examinees received two days of onsite training in administering WISC–IV with Q-interactive, and they conducted several practice administrations before the study began.

Testing took place in San Antonio, TX and Newark, NY (near Rochester). Most administrations were conducted at the child's school, the Pearson office in San Antonio, or a public location (e.g., a church or library). Examiners who were not Pearson employees were paid for their participation. Initially, all administrations were required to be video recorded (with the consent of the examinee's parent), but because many schools prohibit video recording students, the requirement was dropped midway through the study; however, most administrations were video recorded.

Procedure

As each case was scheduled for testing, it was randomly assigned to either the standard or the Q-interactive administration format, with the requirement that the cases within each age-by-gender-by-SES "cell" would be divided equally between the formats. Ethnicity was monitored so that, at each age, half of the examinees in each ethnic group would be assigned to each format. All examiners administered cases using both formats.

Each examinee took the complete WISC–IV in standard subtest sequence, in the assigned format. They then took, in standard (paper) format, the *Kaufman Brief Intelligence Test™, Second Edition* (KBIT™-2; Kaufman, 2004), which yields Verbal and Nonverbal ability scores; the Speed of Information Processing subtest of the *Differential Ability Scales®, Second Edition* (DAS®–II; Elliott, 2007); and the Letter Span subtest of the WISC–IV Integrated.

For all subtests except the Processing Speed subtests, examiners' item scoring decisions were used in analysis (although any errors in calculating subtest raw scores were corrected by Pearson staff). The Q-interactive examiner interface may affect how examiners score items, and so their decisions are an important part of the study. On the other hand, the Processing Speed subtests are scored post-administration in the identical manner for paper and digital formats, so the response booklets for those subtests were rescored by Pearson staff to ensure that there were no scoring errors.

The data was reviewed for quality by inspecting the bivariate scatterplots of scaled scores between pairs of WISC–IV subtests and between those subtests and the covariate tests, as well as inspecting the residual WISC–IV scaled scores (i.e., differences between actual scores and the scores predicted from the regression model based on demographics and the covariate tests). One case was excluded because it was an extreme outlier in the scatterplot between WISC–IV Matrix Reasoning and KBIT-2 Nonverbal, and one case was excluded because of an extreme residual.

Multiple regression was conducted for each WISC–IV subtest, in which the subtest's scaled score was the dependent variable and demographics, the covariate tests, and format were the independent predictor variables. The demographic variables were gender, SES (coded 1 to 5), and dummy codes for the ethnicity groups other than white. Format was also dummy coded (0 = standard administration, 1 = Q-interactive administration). Each analysis included all of the predictor variables. The output of interest was the unstandardized regression weight for format, which is a direct measure of the size of the format effect in WISC–IV scaled score units. The effect size is the unstandardized regression weight divided by 3 (the standard deviation of scaled scores).

Additional analyses were carried out to see whether format effects existed for population subgroups defined by ability level, age, gender, ethnicity, or SES. This was done by comparing the scaled scores from Q-interactive administrations with the expected scaled scores for paper administrations, as predicted by the covariate tests and demographics (gender, ethnicity, and SES). The prediction equations were generated on the basis of the 175 examinees tested with the paper administration. For each Q-interactive examinee, a residual score was calculated by subtracting their predicted scaled score from their obtained score. These residuals represented the effect of the Q-interactive format. Finally, the relationship of these residuals to each population characteristic of interest (ability, age, gender, ethnicity, and SES) was evaluated using linear correlation for the continuous variables (ability, age, and SES), the t test (with unequal variances) for gender, and analysis of variance for ethnicity (African American, Hispanic, white, and other).

Results

Table 2 reports the means and standard deviations of scores on the WISC–IV subtests and the covariate tests for each format and for the sample as a whole. 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–IV subtests and covariate tests by administration format

	Standard Format		Q-interactive		Total Sample	
Subtest or Covariate	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Arithmetic	10.0	2.8	10.3	2.7	10.2	2.8
Block Design	10.3	3.0	10.2	3.2	10.3	3.1
Cancellation	9.8	2.7	9.6	2.7	9.7	2.7
Coding	9.7	2.9	9.8	2.9	9.7	2.9
Comprehension	10.0	2.6	10.0	2.7	10.0	2.6
Digit Span	9.9	2.7	10.2	2.7	10.0	2.7
Information	10.5	2.9	10.7	2.8	10.6	2.9
Letter-Number Sequencing	10.2	2.7	10.7	2.5	10.4	2.6
Matrix Reasoning	10.8	2.6	11.4	2.9	11.1	2.8
Picture Completion	9.3	2.7	9.9	2.5	9.6	2.6
Picture Concepts	10.4	2.5	11.0	2.7	10.7	2.6
Similarities	10.6	3.0	10.7	2.9	10.7	2.9
Symbol Search	10.3	2.6	10.5	2.7	10.4	2.7
Vocabulary	10.2	2.9	10.4	2.9	10.3	2.9
Word Reasoning	10.4	2.8	10.8	2.5	10.6	2.7
KBIT™-2 Verbal	102.0	15.2	102.5	13.2	102.3	14.3
KBIT™-2 Nonverbal	102.4	14.2	101.5	15.5	102.0	14.8
DAS®–II Speed of Info. Proc.	53.0	10.3	54.3	10.1	53.7	10.2
WISC®–IV Integrated LSN	10.4	2.8	10.2	2.9	10.3	2.8
WISC®–IV Integrated LSR	10.3	2.8	10.1	2.6	10.2	2.7
N	175		169		344	

Note: All scores are scaled scores (*M* =10, *SD* = 3) except KBIT™-2 (*M* =100, *SD* =15) and DAS®–II (*M* = 50, *SD* = 10).

Table 3 shows, for each WISC–IV subtest, the multiple correlation with the predictor variables, the unstandardized regression weight for format, the *t* value associated with format as a predictor, and the effect size. With the exception of Cancellation, all subtests had a multiple correlation between .50 and .75 with demographics and the covariate tests. The function of the predictor variables was to account for a portion of subtest score variance and thereby make the analysis of that subtest more powerful by reducing the amount of variance to be explained. So, the variability in multiple correlations indicates that some analyses had greater statistical power than others.

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

Subtest	*R*	Unstandardized Regression Weight	*t*	Effect Size
Arithmetic	.65	0.29	1.25	0.10
Block Design	.59	0.05	0.19	0.02
Cancellation	.39	–0.21	–0.75	–0.07
Coding	.55	0.02	0.09	0.01
Comprehension	.58	0.01	0.05	0.00
Digit Span	.64	0.38	1.63	0.13
Information	.73	0.21	0.96	0.07
Letter-Number Seq.	.56	0.53*	2.24	0.18
Matrix Reasoning	.70	0.80**	3.61	0.27
Picture Comp.	.55	0.58*	2.39	0.19
Picture Concepts	.50	0.63**	2.52	0.21
Similarities	.72	0.05	0.21	0.02
Symbol Search	.57	0.10	0.43	0.03
Vocabulary	.75	0.14	0.68	0.05
Word Reasoning	.67	0.37	1.71	0.12

Note: A positive effect size indicates higher scores with Q-interactive.
* $p < .05$, ** $p < .01$

Two subtests had positive effect sizes that exceeded the pre-established criterion of 0.20: Matrix Reasoning (0.27) and Picture Concepts (0.21). For these two subtests and two others (Letter-Number Sequencing and Picture Completion) there was a statistically significant effect of the Q-interactive administration format, with Q-interactive yielding higher scores. By comparison, on the WAIS–IV, the effect sizes for the common subtests were: Matrix Reasoning, 0.10; Letter-Number Sequencing, –0.04; and Picture Completion, –0.17. Thus, the WISC–IV findings for these subtests were unlike the WAIS–IV results.

A number of steps were taken to investigate possible causes of the format effects on the WISC–IV, focusing in particular on Matrix Reasoning. The first step was to verify the accuracy of subtest scoring on both the Q-interactive and paper administrations. Scoring was found to be accurate for both administration formats.

The next step was to review video recordings of the Q-interactive and paper administrations. These videos were shot from a position above and slightly behind the examiner, so that it was possible to see the examiner's screen (or record form and manual) as well as the examinee's tablet, which lay flat on the table. Along with the audio, this made it possible to see what the examiner and examinee said and did throughout the administration, including what response option the examinee touched and what capture and scoring buttons the examiner touched. Therefore, item scoring could be checked and verified for every case that was reviewed.

Cases that made the greatest contribution to the format effect were prioritized for review. This was done by selecting cases according to the size of the residual, that is, the difference between their subtest scaled score and the scaled score that was predicted from a multiple regression equation that used all predictor variables except format. The first cases to be reviewed were those where the score obtained using the Q-interactive administration was much higher than expected, or the score obtained in a paper administration was much lower than expected. If the cause of the format effect was capable of being observed, then it should definitely have been apparent on these cases. However, no administration, recording, or scoring errors, or examinee behaviors, were detected that would have been capable of accounting for the format effects. A few errors were observed, but they went in both directions and were too infrequent to explain the effect.

An ancillary strategy for identifying a source of the format effects was to calculate the effect for each examiner, on the assumption that the effect might have been caused by some systematic behavior on the part of one or a few examiners. The examiner-specific effect was calculated by subtracting the examiner's average residual with paper administration from the examiner's average residual for Q-interactive administration. Almost all examiners had tested enough examinees with each format to make this calculation possible. The administrations of the examiners with the largest effects were reviewed but, again, these did not reveal any systematic errors or differences in examinee behavior.

In general, examinees appeared to behave in the same way regardless of administration format. One exception was that on Matrix Reasoning they usually touched their answer choice in a Q-interactive administration and said the number of their answer choice in a paper administration.

The possible existence of differential format effects for examinees of different ability levels or different demographic characteristics was investigated in a series of analyses. The measure of format effect was a residual score that was slightly different from the one described previously. This residual was the difference between the actual Q-interactive scaled score and the scaled score that would be expected on a paper administration (as predicted from the covariate tests and demographics, using prediction equations derived from the paper-administration subsample). Results are shown in Table 4. For the continuous variables (ability, age, and SES), the measure of relationship was the correlation coefficient between each variable and the residuals; for gender, it was the *t* statistic; and for ethnicity, it was the *F* statistic from analysis of variance.

Table 4 Relationship of format effect to ability level and demographics

	Correlation				
Subtest	**Ability**	**Age**	**SES**	**Gender *(t)*[a]**	**Ethnicity *(F)***
Arithmetic	.03	.10	.17*	1.85	0.97
Block Design	–.10	.05	–.09	1.35	0.20
Cancellation	–.09	.08	–.00	–0.19	1.39
Coding	–.06	.00	–.05	0.80	1.76
Comprehension	–.09	.08	.06	–0.64	0.49
Digit Span	–.05	–.10	.05	–0.23	2.36
Information	.01	–.13	–.02	–0.19	0.31
Letter–Number Seq.	.01	.08	.05	0.69	0.22
Matrix Reasoning	.00	–.07	.00	0.03	0.43
Picture Completion	–.08	.06	–.08	2.14*	0.47
Picture Concepts	.05	–.16*	.11	2.00*	2.51
Similarities	.00	–.08	.03	0.75	1.93
Symbol Search	–.06	–.08	.10	–1.08	0.31
Vocabulary	–.02	–.08	.13	–1.29	1.66
Word Reasoning	–.09	–.20*	–.02	–0.32	0.96

[a] A positive value of *t* means that the format effect was greater for females.
*p<.05.

No statistically significant correlations were observed between format effect and ability, indicating that the effect of Q-interactive was the same for low-ability and high-ability examinees. On two subtests (Picture Completion and Word Reasoning) there were significant negative correlations with age, meaning that younger examinees benefitted more than older examinees from Q-interactive. On Arithmetic, examinees whose parents had higher levels of education showed a greater Q-interactive format effect than those whose parents had less education. Females benefitted more than males from the Q-interactive format on two subtests (Picture Completion and Picture Concepts), and there were no significant ethnic differences in format effect.

Discussion

Throughout the discussion of possible Q-interactive administration format effects on subtest scores, it is important to keep in mind that the studies completed to date have used nonclinical samples. The potential effects of using the Q-interactive interface with individuals with particular clinical conditions are not yet known. Because digital effects are minimal to non-existent among nonclinical samples, any digital effects that may be observed for a particular clinical group would necessarily be due to an interaction of the digital format with the unique clinical characteristics of that group.

The WISC–IV equivalence study showed very small effect sizes (well below the 0.20 threshold) for eleven of the fifteen subtests. In particular, several subtests that had required special attention in the WAIS–IV study (Information and the Processing Speed subtests) showed negligible effect sizes on the WISC–IV.

Figure 1 is a comparison of the Q-interactive effect sizes of the WISC–IV with those from the prior study of the WAIS–IV. There are a few noteworthy differences between the two sets of results. The first is that only one of the WISC–IV effect sizes was negative, whereas the effect sizes for the WAIS–IV subtests were fairly evenly divided between positive and negative. The WISC–IV effect sizes were (with a few exceptions) very small, but the pattern of directionality is an interesting topic for future investigation. The most obvious difference between the studies, of course, is the age of the examinees. We are seeing a tendency for children and adolescents to do slightly better than adults when digital devices are used to administer a test.

Secondly, whereas on the WAIS–IV the Perceptual Reasoning subtests showed a mixed pattern of (small) effects, on the WISC–IV they displayed consistent positive effects, including two effect sizes above 0.2. Picture Completion is especially interesting because its effect size changed from –.17 on the WAIS–IV to .19 on the WISC–IV. As reported above in the discussion of video review of WISC–IV administrations, there was no visible evidence in examiner or examinee behavior that would explain higher scores on Matrix Reasoning, Picture Concepts, or Picture Completion when using Q-interactive.

What implications do the WISC–IV and WAIS–IV equivalence study findings have for assessment practice? Pearson recommends that practitioners keep in mind that scores on WISC–IV Matrix Reasoning and Picture Concepts will tend to be slightly higher with Q-interactive administration, and use that information during interpretation. However, Pearson does not recommend making numerical adjustments to scores, for two reasons. The first reason is that the format effect sizes are small and (with a few exceptions) not statistically significant, meaning that they may reflect sampling error. For the WAIS–IV, only one subtest (Picture Completion, a supplemental subtest) has a format effect that is statistically significant or reaches the 0.2 effect-size threshold; on this subtest, a Q-interactive administration would be expected to yield a score about one-half scaled score point lower, on average. For the WISC–IV, four subtests have statistically significant format effects, two of which (Matrix Reasoning and Picture Concepts) reach the threshold of a 0.2 effect size; scaled scores on a Q-interactive administration of these subtests would average about one-half to three-quarters of a scaled point higher. All of these effects are well below the subtests' standard errors of measurement.

A second reason is that the format effects on individual subtests have little impact on the index scores. FSIQ scores using Q-interactive and standard administrations are virtually identical on the WAIS–IV (0.1 index score point lower with Q-interactive) and very close on WISC–IV (1.5 point higher with Q-interactive). On the index scores, the WAIS–IV and WISC–IV format effects are: Verbal Comprehension, –1.4 and 0.3 index score points; Perceptual Reasoning, 0.6 and 2.4 points; Working memory, –0.9 and 1.3 points; and Processing Speed, 1.5 and 0.5 points. These are all quite a bit smaller than the standard errors of measurement of the index scores, which range from 2.2 to 5.2 points.

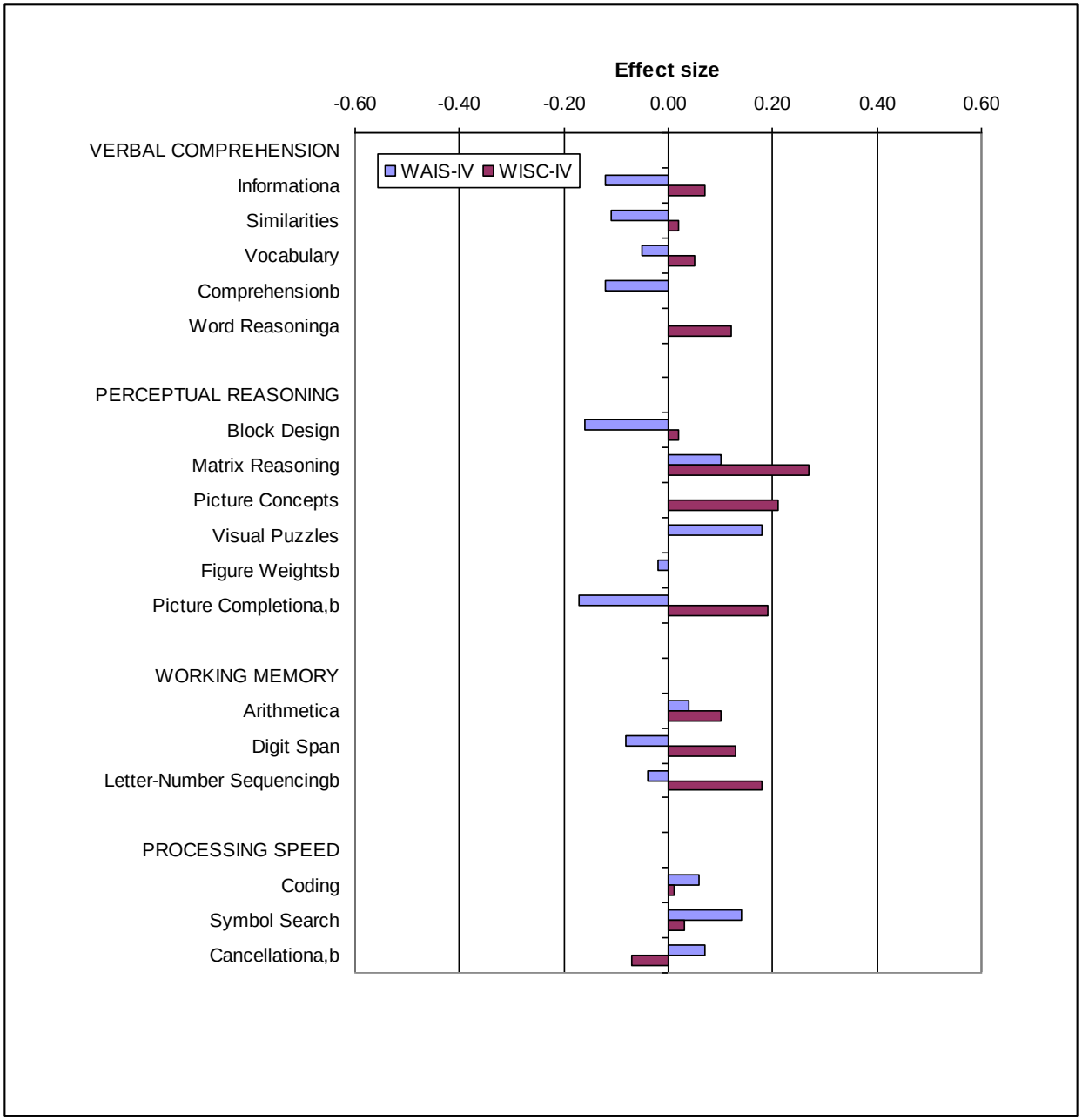

Figure 1: Effect sizes for Q-interactive on WISC®–IV and WAIS®–IV

The finding of only a few significant relationships between format effect and ability level or demographic characteristics allays concerns that Q-interactive might have systematic impact on the performance of subgroups of the population. Of the 75 statistical tests that were performed, only five (7%) produced significant results at the .05 level, little more than would be expected by chance. Given this small number of significant relationships and the fact that they were distributed across the demographic variables, it would be inappropriate to try to interpret them as evidence of a systematic effect.

Children and adolescents tested with WISC–IV exhibited more positive digital effects than the adults tested with WAIS–IV—particularly on subtests that use the examinee tablet—perhaps due to their greater experience with tablets and other touch screen devices. However, there was little indication in either the WISC–IV or the WAIS–IV sample that the format effect was greater for younger than older examinees within each instrument's age range. On the two WISC–IV subtests that showed statistically significant relationships of age with format effect (Picture Concepts and Word Reasoning), the difference in the estimated size of the format effect between age 6 and age 16 was only 1.3 scaled score points. Therefore, there is little evidence to support making digital norms adjustments by age within each instrument.

Finally, the WISC–IV equivalence study adds to the body of evidence about the effects (or lack of effect) of features of interface design on how examinees perform and 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. (2012). Equivalence of Q-interactive administered cognitive tasks: WAIS–IV. *Q-interactive Technical Report 1.* Bloomington, MN: Pearson.

Elliott, C. (2007). *Differential ability scales–second edition.* Bloomington, MN: Pearson.

Kaufman, A. (2004). *Kaufman brief intelligence 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.