Equivalence of Q-interactive® and Paper Administrations of a Speech Sound Task, GFTA™–3 Sounds-in-Words

Q-interactive® Technical Report 10

Ou Zhang, PhD

Shannon Wang, M.A., CCC-SLP

December 2015

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 synced 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.

The focus of this study was the possible effect of digital administration of the Goldman-Fristoe Test of Articulation®–third edition (GFTA™–3; Goldman & Fristoe, 2015) compared to paper administration. Specifically, the examiner-tablet interaction vis-à-vis the digital keypad used to record International Phonetic Alphabet (IPA) characters and its effect on the resulting scores was examined. The digital score and the standard (paper) score for the GFTA–3 Sounds-in-Words test were evaluated. (Note that GFTA–3 refers to Sounds-in-Words as a *test* rather than a *subtest*.) The goal for the GFTA–3 equivalency study was to obtain high inter-rater agreement between standard (paper) and digital administrations. This is in contrast to previous studies that focused on raw score equivalencies. The GFTA–3 data were collected digitally, so if equivalence is demonstrated, then the norms, reliability, and validity information gathered for Q-interactive may be applied to the paper format.

In the initial phase of adapting tests to the Q-interactive system, the goal was to maintain raw-score equivalence between paper and digital administration and scoring formats. In the first two equivalence studies, all 15 *Wechsler Adult Intelligence Scale*®–fourth edition (WAIS®–IV; Wechsler, 2008) subtests and 13 of 15 *Wechsler Intelligence Scale for Children*®–fourth edition (WISC®–IV; Wechsler, 2003) subtests yielded comparable scores in the Q-interactive and 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. The sixth study evaluated all subtests of the *Wechsler Memory Scale*®–fourth edition (WMS®–IV; Wechsler, 2009b), which were found to be equivalent. In seventh study, four tests (Linguistic Concepts, Recalling Sentences, Following Directions, and Formulated Sentences) of the *Clinical Evaluation of Language Fundamentals*®–fifth edition (CELF®–5, Wiig, Semel, & Secord, 2013) all met the equivalence criterion.

In all the equivalence studies, it is assumed that Q-interactive administration may affect test scores for a number of possible reasons, including the following:

- *Examinee interaction with the tablet.* To minimize effects of examinee-tablet interaction that might threaten equivalence, physical manipulatives (e.g., CMS Dot Locations grid) and printed response booklets (e.g., D–KEFS Trail Making) were used with the Q-interactive administration. Although these physical components may be replaced, eventually, by interactive digital interfaces, the degree of adaptation required could cause a lack of raw-score equivalence. More extensive development efforts would then be required to support normative interpretation and provide evidence of reliability and validity. For GFTA–3 administration, effects of examinee-tablet interaction were not a concern because the examinee views stimulus pictures on a screen, but response capture is not dependent upon him or her touching the screen.
- *Global effects of the digital assessment environment.* Global effects go beyond just the examinee's or examiner's interaction with the tablet. For example, a global effect was observed in an early study in which the examiner used a keyboard to capture the examinee's verbal responses. Examinees appeared to slow the pace of their responses so as not to get ahead of the examiner. For GFTA–3 administration, some examiners reported that examinees appeared more engaged with the stimulus pictures depicted on a screen, and the digital pictures held an examinee's attention for a longer period of time. However, regardless of the duration that an examinee looked at a stimulus picture, his or her verbal labeling of the picture did not differ. That is, an examinee responded "house," whether he or she saw the picture on a digital screen or on a printed page in a stimulus book.
- *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.

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 is not a problem. One might say that a reasonable objective for a new technology is to produce results equivalent to those from examiners who use the standard paper format correctly. The digital format should not replicate administration or scoring errors that occur in the standard format. On the other hand, if it appears that a digital effect is due to a reduction in accuracy on the part of either the examinee or the examiner, then the first priority is to modify the Q-interactive system to remove this source of error. Only if that were not possible would the effect be dealt with through norms adjustment.

It is imperative that equivalence studies incorporate a method of checking the accuracy of administration, recording, and scoring in both digital and standard formats. Only in this way can score discrepancies be attributed to one format or the other, or to particular features of either format. All or most of the Q-interactive equivalence study administrations were video recorded to establish the "correct" score for each item and subtest. These recordings had the additional benefit of showing how examiners and examinees interacted with the test materials in each format.

As a whole, the equivalence studies indicate that examinees respond in a similar way when stimuli are presented on a digital tablet rather than a printed booklet. Also, the cumulative evidence shows that when examiners use the kinds of digital interfaces that have so far been studied in place of a record form, administration manual, and stopwatch, they obtain the same results.

Equivalence Study Designs

Several experimental designs have been employed in Q-interactive equivalence studies. In most of them, each examinee takes a subtest only once, in either digital or paper format. This approach avoids any changes in the way an examinee interacts with the task as a result of having done it before. Ideally, a study will detect any effects that the format may have on how the examinee and examiner interact with the task when they encounter it for the first time. Study designs in which there is only a single administration to each examinee provides a realistic testing experience.

One type of single-administration design is the *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 1–2, and 5 and 6.

Another type of single-administration design, called *dual-capture*, was used to capture the data for the GFTA–3 study. The dual-capture design 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. An evidence of interscorer agreement across the two formats indicates whether the format affects the response-capture and scoring processes.

Selection of Participants

The Q-interactive equivalence studies prior to this study have used samples of nonclinical examinees to maintain focus on estimating the presence and size of any effects of the digital format. Because the possible effects of computer-assisted administration on individuals with particular clinical conditions are not known, the inclusion of examinees with various disorders in the sample could obscure the results. However, the GFTA–3 study included examinees with a diagnosis of speech sound disorder as well as examinees who were developing speech normally. Understanding of how examiners record an examinees' misarticulated responses was the focus of this study. Children who demonstrate typical speech sound development produce most speech sounds by age 5, so to provide examiners with opportunities to record misarticulated speech sounds by examinees, the sample included children with a clinical diagnosis of speech sound disorder.

The amount of demographic control required for the sample depends on the type of design. In the equivalent-groups designs, it is important that the samples being compared represent the general population (gender, ethnicity, and socioeconomic status [education level]) and that the two groups are demographically similar to each other. In retest and dual-capture designs, which focus on within-examinee comparisons, examinee characteristics are less significant; however, it is important for the sample to have enough diversity in ability levels and response styles to produce varied responses so that the different features of the digital interface can be evaluated.

The examiners who have participated 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.

GFTA–3 Equivalence Study

Method

Measures

GFTA–3 is an individually administered assessment used to measure speech sound abilities in the area of articulation in children, adolescents, and young adults ages 2 to 21. The Sounds-in-Words test was identified for study by the research team because its Q-interactive examiner interfaces have features that

- could plausibly affect the examiner's ability to capture and score an examinee's responses accurately and
- differ from the test formats already shown to be equivalent to a paper administration in other studies.

The Sounds-in-Words test requires the examiner to transcribe the examinee's oral response accurately and completely, using the symbols from the International Phonetic Alphabet (IPA). Using digital response capture, the examiner uses a keypad that displays each of the IPA symbols that represent the sounds in Standard American English. For paper response capture, the examiner transcribes the examinee's response using IPA symbols. The examiner relies on his or her recollection of the IPA symbols rather than be prompted by a visual cue (i.e., IPA symbols displayed on a digital keypad).

Participants

Ten children, ages 3 years 6 months to 6 years 11 months, were administered the GFTA–3 Sounds-in-Words test during the standardization research phase. The sample had the following composition: 30% female and 70% male, and 80% White and 20% Hispanic race/ethnic origin. The test administrations were recorded. The recordings specifically focused on each child's face so that examiners who would be transcribing the responses for this inter-rater study were able to clearly see the child's mouth movements as well as hear his or her verbal responses.

Eleven examiners who were qualified and experienced in administering and scoring speech/language tests to children and adults participated in the response capture study. Each examiner independently scored 10 protocols.

Procedure

Data collection for this study took place at Pearson's office in San Antonio, Texas between October 31, 2014 and January 10, 2015.

The examiners received onsite training in reviewing videos of the children who were administered the GFTA–3, and capturing the children's responses digitally on Q-interactive and on a paper record form. Each examiner was assigned to view and capture responses for 10 videos of children who were administered the GFTA–3. Each examiner captured responses for five examinees using Q-interactive and five examinees using a paper record form. The examiner viewed and captured responses from the videos in a specified order so that a video was not always the first or last that was viewed. Additionally, Q-interactive and paper record forms were used in a specified order so that neither format was always the first or second that was used.

The analysis of the inter-scorer agreement study focuses on the "change score" for each examiner (i.e., the change in score from the first examiner to other examiners). If there is no effect of format, the expected inter-scorer agreement will be high across the formats. If there is a format effect, the inter-scorer agreement is expected to be low. The inter-scorer reliability coefficients were calculated according to appropriate intraclass correlation procedures (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Total test scores were used in the analysis.

Results

Table 1 reports the characteristics of the sample that took each sequence.

Table 1 Demographic Characteristics of the Sample

	Sample (%)
Gender	
F	30.0
M	70.0
Parent education Level	
1: 0–12years of school, no diploma	10.0
2: High school diploma or equivalent	—
3: Some college or technical school, associate degree	70.0
4: Bachelor's degree or more	20.0
Race/Ethnicity	
Black	—
Asian	—
Hispanic	20.0
Other	—
White	80.0
Region	
Midwest	80.0
Northeast	20.0
South	—
West	—

The interscorer reliability was .92 for the GFTA–3 Sounds-in-Words test for the 11 examiners across both digital and paper formats. This result shows that although these scores were captured through different formats (scores from digitally assisted score capture and paper score capture), they were scored reliably.

Discussion

The interscorer reliability result obtained with the GFTA–3 study indicates that the examiner's scoring pattern is consistent between response capture using different formats. GFTA–3 assesses the same speech sound constructs regardless of the delivery format (i.e., Q-interactive or standard paper). The high inter-score reliability obtained with the scores comparison between Q-interactive and standard paper versions suggests that neither the target construct nor the score capture is altered by the test format; administration, response capture, and scoring is the same for Q-interactive and standard paper.

Taken together, these results provide evidence that the GFTA–3 digital version produces scores that are useful in the assessment of speech sound disorders.

This GFTA–3 equivalence study adds to the body of evidence about the effect (or lack of effect) of features of digital interface design on how examiners capture and score responses, particularly for younger children. It adds new information about the accuracy of automatic scoring of examiner touches. As this body of knowledge grows, it will support generalization to other tests of the same type and features.

References

Q-interactive technical reports:

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

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

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

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

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

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

Daniel, M. H. (2014). *Equivalence of Q-interactive and paper administration of Language Tasks: Selected CELF®–5 Tests. Q-interactive Technical Report 7.* Bloomington, MN: Pearson.

Study References

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

Delis, D., Kaplan, E., & Kramer, J. (2001). *Delis-Kaplan executive function system®.* Bloomington, MN: Pearson.

Delis, D., Kramer, J., Kaplan, E., & Ober, B. (2000). *California verbal learning test®,* second edition. Bloomington, MN: Pearson.

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY®–second edition*. Bloomington, MN: Pearson.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

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

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

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

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

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