Q-interactive: Choosing Sample Sizes for Equivalency Studies

Whitepaper

Mark H. Daniel, PhD
Senior Scientist for Research Innovation

May 2014

Introduction

How were the sample sizes determined for the Q-interactive Equivalency Studies?

Prior to inclusion in the Q-interactive library, each new type of subtest undergoes an equivalency study to evaluate whether scores from Q-interactive testing are interchangeable with those obtained from paper-and-pencil testing. Currently, raw scores from Q-interactive are interpreted using paper-pencil norms, and the equivalency studies support the validity of this practice.

The sample sizes for Q-interactive equivalency studies have been guided by considerations of statistical "power," which is the probability of finding a statistically significant format effect if the true value of the format effect in the population has a particular specified value. At the start of the Q-interactive equivalency research program, the research team selected an effect size of .20 as the smallest effect that would be a threat to the use of Q-interactive results interchangeably with scores from paper-format administration. An effect size of 0.20 would mean that the expected value of an examinee's score was one-fifth of a standard deviation higher (or lower) with Q-interactive than with paper. This is equivalent to 0.6 scaled-score points (where M=10 and SD=3), or 3 standard-score points (M=100, SD=15).

The equivalency studies were designed to provide the maximum power with a given sample size, while still meeting other requirements to ensure validity. It is important to remember that an equivalency study is not a norming study; it is, rather, an experiment designed to measure the format effect in the most direct way possible. This is best accomplished by incorporating experimental and/or statistical controls. Some studies have been able to use two very highly-efficient designs, retest and dual-capture, but these are appropriate only for certain kinds of subtests. The alternative approaches are equivalent-groups designs with either random or nonrandom assignment to groups; these designs are always appropriate, but because they incorporate less internal control, they require substantially larger samples.

Design Types

Retest

This design has a high degree of control, because each examinee serves as his/her own control. Each examinee takes the subtest twice, once in each format. Half of the examinees take the Q-interactive format first and the other half take paper first. We obtain added control by sampling demographically-matched pairs of examinees and randomly assigning one of each pair to each sequence group. Assuming a moderate level of retest reliability (0.80), a sample of 30 examinees (15 matched pairs) provides a high level of power. This design is appropriate when the experience of taking the test the second time is not significantly different from the initial experience. Practice effects are acceptable, but if the examinee's approach to the task is affected by having already taken it, this design probably should not be used.

Dual capture

This design also incorporates a high level of control, but it is appropriate only for subtests where the only difference between Q-interactive and paper administrations is how the examiner captures and scores the examinee's behavior. It uses a small number of video recorded administrations that are scored by a large number of examiners using either the Q-interactive or paper format. Thus, sampling error due to examinees is eliminated. The relevant sample size is the number of independent scorings of the identical administration. The typical Q-interactive equivalency study has used 10 administrations (designed to present a range of performance levels) each of which is scored by 10 examiners, five of whom use paper and five of whom use Q-interactive, yielding 10 scores for each case. Assignment of examiners to cases and formats is carefully balanced to eliminate any relationships among examiners, administrations, and formats. The ultimate analysis is based on the average score difference between 50 Q-interactive scorings and 50 paper scorings of the same administrations.

Equivalent groups

This method has the benefit of matching actual testing practice because the examinee experiences the test only once, in one format or the other. The cost is a relatively low degree of experimental control. The design relies on the equivalence of the two groups, each taking the test in a different format. The most elegant design uses random assignment to groups to ensure that they are equivalent on all characteristics, whether measurable or unmeasurable. Even when covariate tests are included (to provide a degree of statistical control), samples of 150 to 200 cases per group are needed to obtain the desired level of power. An alternative design leverages a test battery's existing norm sample. Each new examinee takes half of the subtests in Q-interactive and half in paper. The paper-administration scores are used to predict performance on the subtests administered using Q-interactive (using a multiple regression equation developed from the norm sample), and the difference between actual and predicted Q-interactive scores is the measure of format effect. This design requires about 100 cases to provide the desired level of statistical power.

Technical Report Publications

Findings from Q-interactive Equivalency studies are published in technical reports, which can be found under the Research section at HelloQ.com/Home.