Paper 2

Taking Short Cuts

Our second presentation of the symposium provided an overview of the challenges that test developers often face when trying to develop shorter assessments that are reliable, valid and fair.

For a long time, much of the research literature and practice in this area has made the assumption that longer tests are more effective, particularly when it comes to applications which involve decision-making, such as selection. This view was discussed throughout this presentation, as well as the difficulties often faced when developing shorter tests.

Length of test and reliability

One of the assumptions of Classical Test Theory is that longer questionnaires provide greater reliability and that greater reliability should lead to increased validity.

However, there is an assumption within the formula (the Spearman Brown Prophecy Formula) that is used to provide an estimate to correct for scale length that each question item within a questionnaire will provide equivalent shared variance and will contribute equally to reliability.

In reality, this is an invalid assumption as some question items are often better than others. To chase a higher internal consistency, longer questionnaires typically contain many redundant items.

Shorter questionnaires typically have a low internal consistency (the question items measure different things). However some well researched and developed questionnaires have been shown to have strong reliability estimates in terms of alternate form and test re-test (Rammstedt et al., 2018, Saville et al, 2012). Arguably, this is the ideal scenario – a test that has broad, short but reliable questionnaire scales.

Length of test and validity

More reliable assessments should lead to assessments of higher validity. However, longer tests are not necessarily more reliable. We looked at what evidence there is to suggest that longer tests are more valid.

An important paper by Burisch (1984) compared the validity of different personality questionnaires and found that length was not an important indicator of validity. More recent work by the same author (Burisch, 1997) indicated that it was possible to select the most valid items from a personality scale. Shorter scales with as little as two items outperformed their much longer counterparts.

So, why don't we simply develop shorter questionnaires?

The first issue is that despite shorter scales being valid with a small number of items, some behavioral traits have a degree of breadth in the content that they measure. This may make it difficult to cover the content with a small number of items.

However, there is a solution to this. Measuring behavioral traits in a scale hierarchy with narrow traits at the bottom, summing to create broader trait measures or factors at the top of your personality scale hierarchy can provide a solution to this issue. For example, a short measure of the Big Five has recently been developed by Soto and John (2017) which, as well as measuring the Big Five, measures three sub-facets using just 30 items.

A second issue for short scales is that they can struggle to have sufficient variance to profile scores. This issue was highlighted by MacIver (1997) who found that five-point Likert formats and most-least formats, which select from three or four response options, typically only provide three points towards the overall scale score.

When profiling scores, on a 1 to 10 scale, at least five and often many more items will be typically needed. To counter this, it is possible to use other formats. For example, the Wave personality questionnaire uses an interactive normative and ipsative format based on a nine-point Likert scale and six alternative items options to be ranked. This delivers a much larger range of responses from one single interactive online questionnaire item.

A final consideration noted for short scales to be used in decision-making is which level in the personality hierarchy is most valid.