Testing non-cognitive attributes in selection centres: how to avoid being reliably wrong

Fiona Patterson and Eamonn Ferguson

In this issue, Gafni et al. provide a strong and timely case for evaluating the reliability and validity of multi-station selection centres (SCs, sometimes referred to as assessment centres). In developing this important research topic, we highlight: (i) important differences between SCs and objective structured clinical examinations (OSCEs); (ii) the importance of demonstrating the construct validity of selection methods; and (iii) psychometric issues surrounding the reliability of observer ratings and the use of generalisability theory (G theory).

First used over 60 years ago, SCs are a measurement methodology based on a multitrait-multimethod approach (MTMM). Although they have some conceptual similarity, SCs differ from OSCEs because, in an OSCE, each station assesses a candidate on one key skill and is usually observed by one assessor. By contrast, in an SC, the same key attribute is assessed across multiple situations (interviews, role plays, simulations, written exercises, etc.) which are observed by a number of trained assessors. Thus, in principle, the SC affords a more reliable assessment as it allows for multiple observations of the same key attributes by multiple observers. In terms of procedural justice, candidates tend to judge SCs as fairer than cognitive ability tests and knowledge tests as, in SCs, candidates have multiple opportunities to perform and the high-fidelity simulations used are perceived as more job-relevant.

When designing a multi-station system, it is essential to: (i) identify the attributes that are key to the target role, and (ii) show that these are assessed using an appropriate methodology. With careful design, the increase in reliability derived from using multiple exercises in an SC should equate to greater validity. A frequent question concerns how many stations are needed to achieve a reliable result. This is something that G theory can address; however, we argue that the primary question should ask how valid a specific selection method is in testing the specific target attribute. Thus, an essential requirement, often overlooked, is to conduct a thorough multi-source, multi-method job analysis to determine the attributes to be tested. Having first determined the target attributes, it is then necessary to select which methods best elicit relevant information and, subsequently, how many testing episodes are needed to test these reliably. The primary focus concerns the target attributes to guide SC design, not the number of stations. Although Gafni et al. refer to the target attributes, they do not fully detail the rationale for choosing them, the relationships among them, or how the method of assessment was chosen to best elicit the attributes being targeted.

Job analysis research has shown that important attribute domains for
multi-station OSCEs because, in an OSCE, each station assesses a candidate on one key skill and is usually observed by one assessor. By contrast, in an SC the same key attribute is assessed across multiple situations (interviews, work simulations, written exercises, etc.) which are observed by a number of trained assessors. Thus, in principle, the SC affords a more reliable assessment as it allows for multiple observations of the same key attributes by multiple observers. In terms of procedural justice, candidates tend to judge SCs as fairer than cognitive ability tests and knowledge tests as, in SCs, candidates have multiple opportunities to perform and the high-fidelity simulations used are perceived as more job-relevant.4,5

Given this distinction, Gafni et al.1 should be commended for examining large-scale, multi-station SCs and highlighting the importance of reliability and validity. In their study, they make use of a range of selection methods including low- fidelity assessments (e.g. a judgement and decision-making questionnaire) and various high-fidelity simulations (e.g. group exercises and patient simulations).1 The use of a combination of low- and high- fidelity methods broadens the range of attributes tested beyond those targeted in a single panel interview, such as is traditionally used in medical selection. As a next stage, a predictive validity study is required to inform the utility of the selection system. For example, a recent validity study of postgraduate selection shows an SC to add incremental validity in predicting job performance over low-fidelity assessments such as situational judgement tests, which target important non-cognitive attributes (e.g. empathy and integrity).7,8

When designing a multi-station system, it is essential to: (i) identify the attributes that are key to the target role, and (ii) show that these are assessed using an appropriate methodology. With careful design, the increase in reliability derived from using multiple exercises in an SC should equate to greater validity. A frequent question concerns how many stations are needed to achieve a reliable result. This is something that G theory can address; however, we argue that the primary question should ask how valid a specific selection method is in testing the specific target attribute. Thus, an essential overlooked, is to conduct a multi-method job analysis to be tested. Having first attributes, it is then necessary to select which methods best elicit relevant information and, subsequently, how many testing episodes are needed to test these reliably. The primary focus concerns the target attributes to guide SC design, not the number of stations. Although Gafni et al.1 refer to the target attributes, they do not fully detail the rationale for choosing them, the relationships among them, or how the method of assessment was chosen to best elicit the attributes being targeted.
Job analysis research has shown that important attribute domains for doctors include empathy, integrity, resilience and communication, amongst others. A detailed job analysis ensures the SC has content validity and the attribute domains targeted are well defined. Without an accurate description of the attributes to be targeted, the content validity of an SC is threatened to the extent that, even with 15 or more stations, assessors could be reliably wrong.

So how does a researcher demonstrate that assessors are assessing the target attributes efficiently and accurately? Research consistently shows that a key determinant of the validity of an SC is ensuring the assessors are trained to a performance standard. For example, assessors are trained to systematically record behaviour and to classify their observations of behaviour into meaningful and relevant categories (i.e. attribute domains identified in a job analysis). Using an MTMM approach, the integration of data relies on the pooling of information on the attributes observed in each exercise. This requires an examination of error variance caused by source of information (observer) and assessment methods, in which, for example, Gafni et al. used G theory to identify sources of error.

In examining the reliability of a multi-station selection system, although classical test theory identifies error variance, it does not distinguish between different sources of error. Generalisability theory identifies facets of the test (the types of tests/items, observers, occasions, setting, dimensions, etc.) as potential sources of error and uses analysis of variance (ANOVA) methods to identify variance components (VCs). Larger VCs indicate that a specific facet offers a greater source of error. The aim is to reduce the specific sources of error and test these with a decision study (D study). The real strength of G theory, therefore, lies in identifying the separate, and potentially modifiable, facets of error. Error variance in the context of observation of performance is, in part, a function of: (i) the nature of the assessment, and (ii) the assessor, as well as the interaction between the two. Researchers often refer to assessor (leniency and taste) and task (difficulty) biases. However, with respect to assessors, there are also various other biases to account for (idiosyncratic bias, ‘halo–horns’ bias, ‘Mum effect’, contrast effects, and the effect of the physical attractiveness of the candidate) and it is crucial to distinguish between these as different sources of assessor error variance. In future, researchers should extend their focus to a wider variety of observation biases, especially as the assessment of clinical and nontechnical skills using observer reports is often described as the reference standard.
Although inter-assessor variability is important, rate–rерate reliability or stability (i.e. when assessors rate the same performance twice) should also be evaluated because the difference between the interassessor reliability and stability coefficients provides an estimate of assessor idiosyncrasy (e.g. preferences for certain behaviours, mood, etc.). If certain attributes have high levels of assessor idiosyncrasy, this indicates that: (i) the way the attribute is defined and assessed needs to be changed, or (ii) assessor training must be improved, or (iii) the assessor may simply need to be excluded from future assessments. It is also crucial to recognise that reliability may not equate to accuracy. An evaluation of accuracy could be used to ‘validate’ the reliability of any assessment via external consensual agreement. For example, videoing a random number of stations would allow for a more finely grained analysis of an ongoing interaction. Although still ‘subjective’, the observers would be: (i) external to the interaction, not participants (thus reducing errors arising from interaction); (ii) able to re-review sections of the interaction to arrive at a consensus, and (iii) able to justify their decisions linked to key frames (which can again be later cross-validated). Thus, in vivo ratings can be validated against in vitro consensually agreed criteria.

Practically, great care must be taken when translating the reliability of any selection system to make decisions about individuals. For example, when cut-off scores are used for selection, good reliability among observers does not rule out systematic observer bias. One assessor may rate a group of six candidates on an attribute and give respective ratings of 4, 5, 7, 7, 3 and 8, and another assessor may give the same candidates ratings of 6, 7, 9, 9, 5 and 10, respectively. Although these are perfectly correlated, the second assessor, in absolute terms, is systematically more lenient (mean rating: 7.6) commentaries 241 ª Blackwell Publishing Ltd 2012. MEDICAL EDUCATION 2012; 46: 238–244 241 than the first assessor (mean rating: 5.6). We do not know which of these assessors is more accurate. If the cut-off for the attribute is 5, all candidates pass except the fifth, who fails with an average of 4, although the second assessor gave him a rating of 5. This illustration shows that an indication of the accuracy (not just reliability) of selection is required to allow researchers to demonstrate the extent to which assessments on a particular station are influenced by idiosyncratic and leniency biases.

Gafni et al.1 provide an important contribution to understanding reliability in multi-station selection systems. Generalisability theory is important for validating multi-observer assessments, especially for identifying facet-specific error variance. By extension, we propose that specific aspects of observer bias should be investigated as detailed above. Here, multi-level modeling techniques can account for the hierarchical nature of these data, within- and between assessor effects, lack of independence in the data and potential problems of auto-correlation.18,19 Most importantly, the rationale and justification for why certain attributes are targeted in selection, and matching these to an appropriate selection method, require greater exposition in the research literature.
REFERENCES


