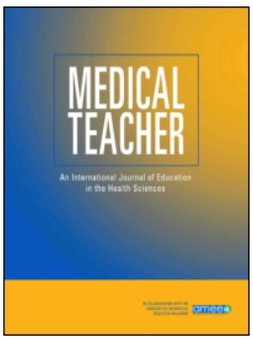
2016-03-03

Twelve tips for assessment psychometrics

Coombes, L

10.3109/0142159x.2015.1060306
Medical Teacher
Informa UK Limited



Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

Twelve tips for assessment psychometrics

Lee Coombes, Martin Roberts, Daniel Zahra & Steven Burr

To cite this article: Lee Coombes, Martin Roberts, Daniel Zahra & Steven Burr (2015): Twelve tips for assessment psychometrics, Medical Teacher, DOI: 10.3109/0142159X.2015.1060306

To link to this article: http://dx.doi.org/10.3109/0142159X.2015.1060306

Published online: 16 Oct 2015.

2015, 1–5, Early Online MEDICAL TEACHER

TWELVE TIPS

Twelve tips for assessment psychometrics

LEE COOMBES, MARTIN ROBERTS, DANIEL ZAHRA & STEVEN BURR
Plymouth University Peninsula Schools of Medicine and Dentistry, UK

Abstract

It is incumbent on medical schools to show, both to regulatory bodies and to the public at large, that their graduating students are "fit for purpose" as tomorrow's doctors. Since students graduate by virtue of passing assessments, it is vital that schools quality assure their assessment procedures, standards, and outcomes. An important part of this quality assurance process is the appropriate use of psychometric analyses. This begins with development of an empowering, evidence-based culture in which assessment validity can be demonstrated. Preparation prior to an assessment requires the establishment of appropriate rules, test blueprinting and standard setting. When an assessment has been completed, the reporting of test results should consider reliability, assessor, demographic, and long-term analyses across multiple levels, in an integrated way to ensure the information conveyed to all stakeholders is meaningful.

Introduction

Assessment psychometrics is the measurement, analysis, and interpretation of performance across qualitative and quantitative assessment, using the best available evidence to provide appropriate and defensible standards. When a student graduates, it is because they are deemed to have acquired the appropriate skills, knowledge, and professionalism required for the next stage in their career as a healthcare professional. Those responsible for approving student progression and final award rely on the support of assessment psychometricians to fairly and precisely process and interpret student assessment data; to state whether a student has met the required standard, and therefore demonstrated the requisite abilities to progress. The quality assurance of standards is necessary to reassure stakeholders: ensuring that assessment decisions are fair for students whilst also maintaining both public safety and the reputations of the awarding institution and the profession.

For any psychometric analysis, there will be a range of individuals with an interest in the outcome, all of whom will have different experience and consider different information when reviewing the validity and reliability of an assessment. The role of producing assessment psychometrics can fall to people from a wide range of backgrounds, each with a contribution to make in providing a reliable and defensible psychometric service. While the obvious choice may be a statistician, those with training in human, healthcare, or social sciences and similar academic backgrounds are likely to provide a different and distinct, but equally valid approach to psychometric analysis. A background in computing, where strong numeracy skills are essential, can be useful for creating bespoke solutions for analysis and feedback. Whoever fulfils the role of psychometrician, there are some basic considerations when quality assuring the development, pre-test preparations, and post-test reporting of assessments.

Development

Tip 1

Nurture and support a quality assurance culture

The culture in which analysis takes place is critical to maintaining stakeholder confidence. Psychometric analysis should be valued by those required to use it, with clear benefits for staff and students built on collaborative relationships between psychometricians and members of the faculty. When mistakes happen, it is important to act professionally in correcting errors to ensure that assessments remain fair. Mistakes may be revealed by psychometric analysis or feedback from other stakeholders. Such errors can be minimized by involving psychometricians throughout the design and development of assessments. Keeping an adverse incidents log can also be helpful, and can support new members of staff in avoiding mistakes that have occurred in the past. By nurturing a culture that encourages honesty and openness, if errors occur they can be admitted and addressed without fear of disclosure leading to recrimination (National Audit Office 2014).

Tip 2

Take an evidence-based approach

Evidence-based practice is not limited to clinical sciences (Hjørland 2011) and is equally important for psychometric

Correspondence: Lee Coombes, Collaboration for the Advancement of Medical Education Research and Assessment (CAMERA), Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK. E-mail: lee.coombes@plymouth.ac.uk

ISSN 0142-159X print/ISSN 1466-187X online/15/0000001–5 © 2015 Taylor & Francis.
DOI: 10.3109/0142159X.2015.1060306

analysis and support. There is often no gold standard, although if one existed for any aspect of analysis then we would not see new ideas being presented, or the evolution of older ideas through necessity or curiosity. Instead there are a range of viable options, each with its advantages, disadvantages and appropriate contexts, but whichever is chosen it should be well documented and supported by evidence from reliable sources. As that evidence base changes over time, the means and methods of assessment may also change and this requires a flexible analytical approach, capable of adaptation when new evidence comes to light. The evidence for change may come from external sources such as the academic literature, from researching and modelling new approaches alongside routine analysis, or seeking the opinions of teachers, administrators, students, or other stakeholders on potential changes. Having a diverse team involved with psychometric analysis can bring expertise, ideas and evidence from different disciplines.

Tip 3

Demonstrate validity

Fundamentally, psychometric analysis aims to quality assure the reasonableness of the interpretation of assessment outcomes and is a major source of validity evidence. Thus providing good metrics are central to ensuring assessment validity. Get it right and we have the foundations for defensible and acceptable assessment, but get it wrong and an assessment can be irreversibly damaged. Downing (2003) provides a wide-ranging list of sources of validity evidence in medical education, each of which can contribute to increasing overall validity. Assessments should be carefully planned, with the gathering of validity evidence being fundamental to test design. This can range from simple face validity where the look and the feel of an assessment should be acceptable to all those involved, through to complex measures of criterion and content validity that show statistically that the aims of the assessment are being met.

Pre-test preparation

Tip 4

Know the rules and regulations

Each institution has its own way of working, and subsequently its own set of rules to adhere to. These may take the form of student handbooks, school, and institutional regulatory manuals, or national quality assurance and legal requirements (Quality Assurance Agency 2013). The people who are most motivated to scrutinise the rules are likely to be those who receive an unsatisfactory outcome from an assessment. It is easy to understand why so many appeals are based on the rules being broken when, as a result of their assessment outcomes, we require a student to do something that has life-changing potential such as repeat part of, or withdraw from, a programme.

Rules should be clearly defined, easily accessible, and continually reviewed (Ricketts & Bligh 2011). School programme and institutional policy administrators should drive this process as they will have in-depth knowledge not only of the rules but also their application. There should also be clear policies on data checking, security and confidentiality. Beyond the written rules, there are some aspects of psychometric analysis that may seem small but could be critical. How many decimal places do you work with and report? Where numbers are rounded which method is used? What happens when a student has extenuating circumstances (ECs)? Where someone has multiple validated ECs, when do we have enough assessment data to make an informed decision on their progress? And how do we assess them fairly against standards or others in their cohort? Sometimes it is easy to spot something that is not explicitly stated in the rules, such as how to treat ECs, and document a contingency plan for when the situation arises. Occasionally a situation presents itself for which there are no rules or guidance, and defensible rules have to be created. In these cases, it is particularly important to document the issue, the available options, which option was chosen to resolve the situation, and the reasons why it was preferred for future reference.

Tip 5

Blueprint to curricula content

Analysis of test results in relation to blueprinting should be part of the psychometric analysis of any assessment, and can provide evidence of validity. Typically we want to have evidence that someone has achieved a required standard, but we can only know what this standard is when we have explicitly detailed learning outcomes. We then need to decide on a mode of assessment that will give us the evidence we require to show each outcome has been achieved.

The blueprint itself can be extremely broad or very narrow in scope, at the level of the programme or its subdivisions, can be uni- or multi-dimensional, and this will be dependent on the purpose and context of the test. The UK General Medical Council's Professional and Linguistic Assessment Board blueprint (http://www.gmc-uk.org/doctors/plab/Blueprint.asp) is a good example of a detailed medical blueprint, aligned to the knowledge expected of a doctor entering the second year of foundation practice in the UK. Each topic, presentation, and condition is listed in the first dimension while the second allows test items to be aligned to the outcomes of Good Medical Practice (GMC 2013). When a test is created, items can be mapped across the blueprint to ensure a balanced sample from all the major and minor areas of the syllabus.

Tip 6

Determine appropriate standard setting methods

The method for setting a standard should be decided proactively and communicated to all stakeholders ahead of the assessment rather than being based on the performance of the assessment. The preferred method might require some prerequisite conditions to be met. For instance, if we choose a

method that requires assessors to consider awarding a new grade this might require changes to published rules, training, mark sheets, student rubrics, data storage systems, report formats, and feedback methods before implementation.

Practically, standard setting methods can fall into two categories. Staples such as Angoff's or Ebel's methods allow a standard to be set once the material in a test has been finalized, as they involve expert judges allocating scores to items from which a standard can be calculated. Alternately, methods such as Borderline Regression rely on calculations based on the assessment outcomes. There are also methods such as those proposed by Hofstee and Cohen-Schotanus that sit between the two, relying on some information recorded from experts ahead of an assessment, but with the cut score being data dependent (Cohen-Schotanus & van der Vleuten 2010; McKinley & Norcini 2014). Whichever approach is taken, the method must be feasible, academically and legally defensible, and otherwise fit for purpose.

It is also prudent to consider the options for when a method fails and note ahead of the assessment under what circumstances any alternative might be applied. We should also specify, prior to running the assessment, any adjustments to be made to the standard once the test has concluded. It is not uncommon to see the standard error of measurement (SEM) used to provide reliable decisions by compensating for the error between a true and an observed score. The exact way in which the SEM is applied will be dependent on the context in which it is used, and the confidence we want in our outcomes (McManus 2012).

Post-test reporting

Tip 7

Understand the calculation, interpretation, and limitations of reliability coefficients

There are many ways to express the reliability of an assessment, and it is a key factor in assuring quality and providing validity (Norman 2014). If the evidence for either is inadequate, then an assessment cannot be used to measure and subsequently make decisions about performance. It is vital that we know exactly how a reliability statistic is reached. Often reliability of an assessment is captured as a single number such as KR20/Cronbach's alpha (Kuder & Richardson 1937; Cronbach 1951) but this may not be appropriate and relying on it for interpreting overall statistical reliability could actually mask vital information about assessment performance. Where the coefficient is suitably high to assure the casual observer that the test is reliable, the truth may be that similar parts of the material have been left unanswered by candidates, inflating the internal consistency. Where it is low, it may be that a test is multidimensional and examinees have different experiences and knowledge, but this does not suggest the assessment is not capable of accurately sampling attainment across a domain. The misunderstood nature of these coefficients mean that without a deeper appreciation of what a statistic is actually telling us, we are in danger of reducing a rich information source down to a single figure and then misinterpreting the information it can provide (Sijtsma 2009).

The best approach to appraising reliability can be to consider multiple measures in context. Statistics based on internal consistency, test–retest, split half, parallel forms, and inter-rater agreement can all add information about a test's performance and its reliability. Generalizability theory can be used to provide coefficients of reliability, with decision studies capable of modelling the impact of potential changes to an assessment (Crossley et al. 2002). Ultimately, we need to demonstrate decisions based on a test are defensible, and we can achieve this by providing a range of information and being confident in its interpretation (Hays et al. 2015).

Tip 8

Produce detailed assessor analysis

Many assessments rely on expert judges to set standards, award grades, and provide scores in a manner that is fair. Before becoming involved in the assessment process, they may be benchmarked or otherwise tested and vetted for their suitability, but when they become part of the assessment process they need to be accountable. What makes this particularly challenging is that occasionally an assessor may appear to be an outlier in the data because all the work they have marked is of a similar standard attracting a limited range of scores. This can be particularly evident when they are only assessing a small number of students. An outlier may only be an outlier in the data set we are examining, so we need to be careful with how we interpret the analysis we have.

In order to ensure the experts and assessors are capable of the task assigned them, detailed analysis of performance should be used. This can highlight outlying or inappropriate performance as they would do with an examinee, but without the ability to judge whether marking is appropriate. Simple measures of variation such as a standard deviation or interquartile range for each assessor can highlight unusually varied or overly consistent marking when compared with others. If more than one person is assessing the same thing, inter-rater agreement statistics such as intra-class correlation coefficients or kappa-type statistics can be used (Shrout 1998). When a range of information has been gathered, the prospect of excluding an assessor or adjusting ratings can be considered.

Tip 9

Produce detailed analysis at multiple levels

Making defensible assessment decisions requires the most accurate information to be available, so the psychometrician needs to capture and summarize data across many levels. At the level of the individual components of the assessment there are simple summaries; such as item difficulty, discrimination, response patterns, item characteristic curves, item information functions, and differential item functions (Livingstone 2006). These can show how each item within an assessment contributes towards the outcome and can also reveal patterns

of behavior, where examinees opt for an answer or attempt a skill that demonstrates a deep understanding of a topic, or is completely inappropriate.

Moving up a level, analysis of grouped or themed assessment items using classical or modern approaches to test theory can help to validate a test by demonstrating that learning outcomes are being met, and also inform and refine the test blueprint. Analysis of the test as a whole should also go a long way to supporting test validity if the correct steps have been followed when creating the assessment. Of course, it is rare that a single assessment is able to completely capture the ability of an examinee, and often examinations are combined to make decisions, so analysis should also take in these broader levels. This can be in the form of analysis across several tests of a single longitudinal assessment such as a progress test (Coombes et al. 2010), or diverse assessments that combine at a modular level to make decisions about progression to more senior study. Standardized scores can facilitate analysis across different test formats.

There are many other ways we can slice data to review an assessment, and providing information that is succinct yet captures everything of importance is always challenging because stakeholders will each view an assessment distinctly and place importance on different aspects of it. To ensure that everyone has the information they need, detailed analysis can be carried out but presented alongside a summary document that contains the key information and acts as a contents table for the detailed sections.

Tip 10

Identify and analyze key demographics

Variations in performance based on gender, ethnicity, and disability status should always be reviewed, but other factors can also be included should they be deemed important. For example, a clinical exam run across several days might include exam time to ensure that there has not been any order bias. If an examination includes assessors, analysis can indicate if marking has been fair across demographic groups of assessors and students regardless of their characteristics.

Demographic analysis should be robust and meaningful, but often it is a case of choosing one or the other. Each group taking a test is likely to have its own characteristics and often a seemingly reliable analysis can be meaningless. Consider the case of ethnicity, where group sizes may dictate which way we produce an analysis. The most detailed version of an ethnicity analysis would include every possible classification, but some groups might naturally have small numbers. While meaningful, this would be unreliable. Grouping ethnicities together may alleviate the problem, but groups can still have small membership and should we find a significant difference between them it is difficult to know how to interpret this finding. Simple information can often be the most useful. While testing for significant differences between groups is always encouraged, simple descriptive statistics can highlight the same issues as extensive significance testing and should not be overlooked.

Tip 11

Make good use of historic data

If everything we do is to be evidence based, historic data provide a source of evidence that can be used to improve our assessment as part of a continual review of our practices and is essential when modelling and examining the impact of potential changes. Where an assessment has evolved over time, data may be weighted so more recent data can be more influential on predictions, or older data removed completely. We often do not know what information might come in useful later, so knowing exactly what data to record and store can be educated guesswork until we have the benefit of hindsight. When variables and files are clearly named so that those with no familiarity of an assessment's history can utilise it, historic data can be the most accessible source of evidence we have available.

Tip 12

Tailor feedback to your audience

Whatever psychometric analyses we carry out, the feedback we create must be meaningful and provide a foundation for change. This may be a student changing their exam or revision strategy, an assessor changing the way they mark, a tutor revising the content of an assessment or an administrator changing the logistics of a test. It is not possible to create a single analysis that everyone will find useful, so recognizing what type of information each person or group requires can help dictate which analysis we carry out. All of our stakeholders are important, with some making life-changing decisions based on our analysis. Not everyone is an expert, so the level of analysis we provide needs to be tuned to the group we are providing feedback for. If the most appropriate analysis is one that few people know or understand, reporting outcomes alongside better known and more accepted statistics can satisfy both psychometricians and feedback users. If they result in different conclusions, a choice needs to be made. It is, therefore, important to be cognizant of the strengths and weaknesses of all potential analytical approaches and be prepared to give expert advice on their interpretation. Providing meaningful feedback is also the last part of our feedback loop, where the feedback we create must be adequate to inform decision making (Coombes et al. 2010; Burr et al. 2013). Creating this loop ensures that there is a clear and open path to change an assessment in light of new evidence provided by modelling and other analyses.

Conclusions

A single statistic is never the whole picture so providing a range of psychometric information allows defensible decisions to be made. Correctly applied psychometrics play a key role in creating a defensible programme of assessment, and should be central to validating and quality assuring defensible assessment decisions. Even when a range of statistical information is available, sometimes this can only act as a signpost for further qualitative investigation. Feedback should be used to improve

all aspects of assessment, and the culture in which it is used must be receptive to the challenges that can come with increased scrutiny and provision of best available evidence.

Notes on contributors

LEE COOMBES, PhD, is a Lecturer in Clinical Education (Assessment Psychometrics) at Plymouth University Peninsula Schools of Medicine and Dentistry.

MARTIN ROBERTS, BA, Cert Ed, MSc, is a Senior Psychometrician at Plymouth University Peninsula Schools of Medicine and Dentistry.

DANIAL ZAHRA, PhD, is a Senior Psychometrician at Plymouth University Peninsula Schools of Medicine and Dentistry.

STEVEN BURR, PhD, is an Associate Professor in Physiology and Deputy Director of Assessment at Plymouth University Peninsula Schools of Medicine and Dentistry.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

References

Burr SA, Brodier E, Wilkinson S. 2013. Delivery and use of individualised feedback in large class medical teaching. BMC Med Ed 13:63. doi: 10.1186/1472-6920-13-63.

Cohen-Schotanus J, van der Vleuten CP. 2010. A standard setting method with the best performing students as point of reference: Practical and affordable. Med Teach 32(2):154–160.

Coombes L, Ricketts C, Freeman A, Stratford J. 2010. Beyond assessment: Feedback for individuals and institutions based on the progress test. Med Teach 32(6):486–490.

Cronbach LJ. 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334.

Crossley J, Davies H, Humphris G, Jolly B. 2002. Generalisability: A key to unlock professional assessment. Med Educ 36(10):972–978.

Downing SM. 2003. Validity: On the meaningful interpretation of assessment data. Med Educ 37:830–837.

General Medical Council. 2013. Good Medical Practice. [Accessed 24 December 2014] Available from http://www.gmc-uk.org/gmp.

Hays RB, Hamlin G, Crane L. 2015. Twelve tips for increasing the defensibility of assessment decisions. Med Teach 37(5):433–436.

Hjørland B. 2011. Evidence based practice. An analysis based on the philosophy of science. J Am Soc Inf Sci Tec 62(7):1301–1310.

Kuder GF, Richardson MW. 1937. The theory of estimation of test reliability. Psychometrika 2:151–160.

Livingstone SA. 2006. Item analysis. In: Downing SM, Haladyna TM, editors. Handbook of test development. London: Routledge.

McKinley DW, Norcini JJ. 2014. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach 36(2):97–110.

McManus IC. 2012. The misinterpretation of the standard error of measurement in medical education: A primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. Med Teach 34(7):569–576.

National Audit Office. 2014. Making a whistleblowing policy work. London: National Audit Office.

Norman G. 2014. When I say… reliability. Med Educ 48(10):946–947.

Quality Assurance Agency. 2013. The UK Quality Code for Higher Education. [Accessed 24 December 2014] Available from http://www.qaa.ac.uk/assuring-standards-and-quality/the-quality-code.

Ricketts C, Bligh J. 2011. Developing a "Frequent Look and Rapid Remediation" Assessment system for a new medical school. Acad Med 86(1):67–71.

Shrout PE. 1998. Measurement reliability and agreement in psychiatry. Stat Methods Med Res 7(3):301–317.

Sijtsma K. 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 749(1):107–120.