Standards for competency assessment measures[1]

1 Introduction

This paper sets out standards that measures for evaluating the competencies trainee PWPs should meet. It recommends that only measures meeting minimum standards should be used by PWP courses.

2 Assessment standards – reliability and validity

Scales for assessment of PWP competence, as all assessment measures, need to meet psychometric standards of reliability and validity. There are different types of reliability and validity and the ways that reliability and validity are classified are not uniform. For measures of PWP competence, the following are the most relevant:

1. Content validity
2. Inter-rater reliability
3. Consistency across different forms/stimuli/situations
4. Convergent validity
5. Criterion/predictive validity

These are discussed in turn below.

3 Content validity

Content validity refers to the extent to which the measure is made up of items which accurately reflect the construct being measured. A measure will be less valid if it contains items which (1) only measure part of the construct (2) reflect a different construct or (3) are inaccurate in how they reflect the construct. Specifically for PWP measures of competence, the key areas are:

- Does the measure contain items that cover all the learning outcomes specified in the national curriculum that it is designed to cover?
- Do the items designed to evaluate a specific learning outcome, accurately and appropriately capture the learning outcome (e.g. would an item which counted the number of times the trainee uttered ‘mh-hm’ be considered to accurately and appropriately capture competence in therapeutic empathy/alliance)?

[1] Document written by a working group convened by University College London under the aegis of the national IAPT team/ NHS England/Department of Health. Originally included as appendix 5 of the PWP Training Review.

- Do the weightings in the marks of the measure accurately reflect the weightings of the importance of the different learning outcomes (e.g. if the learning outcome in module 1 of competence in using common factors to engage patients is considered twice as important as the learning outcome of competence in recognising patterns of symptoms consistent with diagnostic categories, does the marking system reflect this)?
- Are the wordings of the items and scoring criteria clear to markers? Is there a scoring manual?
- Is there a clear pass/fail threshold mark or set of marks that reflects the minimum threshold of acceptable competence for PWP practice?

4 Inter-rater reliability

Inter-rater reliability is the extent to which different raters/assessors give the same score/mark on the measure. For competence assessment scales, this is the extent to which different assessors/examiners watching or listening to the recording of a trainee's performance give the same set of marks. Kappa (percentage agreement corrected for chance) is the standard statistic for assessing inter-rater reliability. Double marking and averaging the scores of the markers improves scale reliability.

5 Consistency across different forms/stimuli/situations

Where there are parallel forms of a scale or a scale is used to measure performance across parallel situations, there will be variation in individuals' performance between scales/situations. The extent of this variation or inconsistency, represents unreliability in the scale as a measure of the construct. This is the case for PWP competency assessments as a PWPs performance will vary between different simulated patients or between different real patients. In testing for this, the variation obtained in PWP performance will technically be both test-retest variation (PWP differences in performance that would occur to exactly the same simulated or real patient if assessed at two different time points) and variation across different forms/stimuli/situations. So pragmatically these are combined in this standard.

6 Convergent validity

Convergent validity refers to the extent to which a measure comes up with similar results to other measures of the same construct. So to what extent it ranks individuals in the same order as other measures. So in validating a new measure of trainee PWP competency one would want to know how it compared in scoring trainees with the current Reach Out scales and how it compared to supervisor ratings of trainee competence. One would not expect the new and old scales to give identical rankings (if a new scale is thought to be more reliable and valid on some of the other validity/ reliability criteria its scores would be expected to differ from the previous scale); but one would expect there to be a reasonable correlation.

7 Criterion/predictive validity

Predictive/criterion validity is the extent to which the measure predicts performance on some outcome that is of relevance to the construct measured. For PWP competence measures, relevant predicted criteria would include satisfaction of patients with their treatment by the trainee and patient clinical outcomes. These would be assessed with patients on the trainees' routine caseloads. As can be seen from this example, one would expect the criterion variable (clinical outcomes and patient satisfaction) to be influenced by multiple factors beyond the construct (trainee competence) so the association between competence measure and criteria would generally be low.

8 Minimum and ideal standards for competence assessment measures

Measures of evaluation of PWP competency used by PWP courses should meet the following minimum standards:

1. Measures should have satisfactory content validity in relation to:
    - Covering all relevant learning outcomes of the national curriculum for the module assessed
    - With items accurately and appropriately assessing each learning outcome
    - Weighted in accordance with agreed weightings of importance of each learning outcome
    - With clear wordings and scoring criteria (including a scoring manual)
    - With agreed pass/fail mark or set of marks
2. Measures should have adequate inter-rater reliability

Ideally measures of trainee PWP competency will also meet the following standards:

3. Measures should demonstrate consistency across the range of stimuli/situations used to measure performance (OSCIs and/or real patients)
4. Measures should demonstrate convergent validity with other previous competency assessment measures and with supervisor competency ratings

Evaluation of predictive validity is not a standard expected of a PWP competency assessment measure used by training courses, but is a relevant and useful issue for courses to research.