A critical appraisal of instruments to measure outcomes of interprofessional education

Matthew Oates & Megan Davidson

CONTEXT Interprofessional education (IPE) is believed to prepare health professional graduates for successful collaborative practice. A range of instruments have been developed to measure the outcomes of IPE. An understanding of the psychometric properties of these instruments is important if they are to be used to measure the effectiveness of IPE.

OBJECTIVES This review set out to identify instruments available to measure outcomes of IPE and collaborative practice in pre-qualification health professional students and to critically appraise the psychometric properties of validity, responsiveness and reliability against contemporary standards for instrument design.

METHODS Instruments were selected from a pool of extant instruments and subjected to critical appraisal to determine whether they satisfied inclusion criteria. The qualitative and psychometric attributes of the included instruments were appraised using a checklist developed for this review.

RESULTS Nine instruments were critically appraised, including the widely adopted Readiness for Interprofessional Learning Scale (RIPLS) and the Interdisciplinary Education Perception Scale (IEPS). Validity evidence for instruments was predominantly based on test content and internal structure. Ceiling effects and lack of scale width contribute to the inability of some instruments to detect change in variables of interest. Limited reliability data were reported for two instruments. Scale development and scoring protocols were generally reported by instrument developers, but the inconsistent application of scoring protocols for some instruments was apparent.

CONCLUSIONS A number of instruments have been developed to measure outcomes of IPE in pre-qualification health professional students. Based on reported validity evidence and reliability data, the psychometric integrity of these instruments is limited. The theoretical test construction paradigm on which instruments have been developed may be contributing to the failure of some instruments to detect change in variables of interest following an IPE intervention. These limitations should be considered in any future research on instrument design.
INTRODUCTION

The World Health Organization recognises the important role of interprofessional education (IPE) in developing a collaborative practice-ready workforce. The Centre for the Advancement of Interprofessional Education defines IPE as occurring ‘when two or more professions learn with, from and about each other to improve collaboration and the quality of care’. A lack of rigorous evaluation of the effectiveness of IPE curricula makes it difficult for educators to confirm their assumptions that students are prepared for the collaborative practice workplace. A range of instruments have been developed to measure outcomes of IPE and collaborative practice. However, in order to provide credible data with which to test the effectiveness of IPE, it is imperative that instruments are developed with robust psychometric properties (validity, reliability). Appropriate scale and score construction enables test scores to discriminate between test respondents and facilitates valid interpretations of these scores.

Although previous work has sought to identify available instruments, there has been no systematic and critical appraisal of their psychometric properties. An inventory of instruments published by the Canadian Interprofessional Health Collaborative (CIHC) provides a qualitative description of each instrument and reports available psychometric data as indicated by the original instrument developer(s). A review by Thannhauser et al. added to the established pool of instruments. To be selected for critical appraisal, the instrument, including its items and scoring system, had to be available. Where a full version of the instrument was not published in a peer-reviewed article, attempts were made to source this through alternative avenues, such as by direct contact with the author(s). The instrument was required to be designed for use with pre-qualification health professional students in learning situations involving more than two professional groups and to be intended to measure one or more IPE outcomes based on Barr et al.’s six-level framework of educational outcomes (Table 1). Barr et al. modified the four-level typology of educational outcomes developed by Kirkpatrick, in order, they argue, to capture the breadth of potential outcomes of well-designed and delivered IPE. At least one peer-reviewed publication reporting the instrument’s development and reliability or validity had to be available.

METHODS

A pool of extant instruments was established based on the CIHC inventory of instruments and a smaller review of a limited number of instruments by Thannhauser et al. It was noted that all instruments identified in the review by Thannhauser et al. had been identified by the CIHC in its inventory. Rather than conduct a full independent search for available instruments, the authors of the present review decided to update the search for the period from January 2010 to June 2014 to identify any new instruments not previously identified. The reason for this was two-fold: (i) the search strategy employed by the CIHC was thorough and the authors of this review were confident the inventory had captured all available instruments, and (ii) given the recent publication of the CIHC inventory at the time of the commencement of this review, it appeared unnecessary to duplicate the work of the CHC. It is worth noting, however, that some minor modification to the original CIHC search strategy (mainly in truncation convention) was necessary to take into account variations in database hosting platforms between institutions. Search results from all included databases were exported to EndNote™ and duplicates were excluded.

The authors independently reviewed titles and abstracts obtained through this search to identify those articles that described a quantitative tool for measuring outcomes of IPE or collaborative practice. The reviewers compared the titles and abstracts included and excluded, and resolved any disagreement through discussion. Full-text articles were retrieved for included citations. Again, the authors independently reviewed the full-text articles. Instruments that had not been previously identified were added to the established pool of instruments.

To be selected for critical appraisal, the instrument, including its items and scoring system, had to be available. Where a full version of the instrument was not published in a peer-reviewed article, attempts were made to source this through alternative avenues, such as by direct contact with the author(s). The instrument was required to be designed for use with pre-qualification health professional students in learning situations involving more than two professional groups and to be intended to measure one or more IPE outcomes based on Barr et al.’s six-level framework of educational outcomes (Table 1). Barr et al. modified the four-level typology of educational outcomes developed by Kirkpatrick, in order, they argue, to capture the breadth of potential outcomes of well-designed and delivered IPE. At least one peer-reviewed publication reporting the instrument’s development and reliability or validity had to be available.

All published studies of the included instruments were sourced via a search of databases (MEDLINE,
Data extraction from all included articles used the Quality Appraisal of Interprofessional Learning Scales (QuAILS), a standardised checklist specifically developed for the purpose of this review. (A full description and user guide are available online in Appendix S1.) The two-part QuAILS checklist was developed on the basis of the Standards for Educational and Psychological Testing (the Standards) and other checklists developed for the purpose of evaluating other measurement scales. Part A considers the qualitative features of the instrument, including a description of the setting in which the instrument was developed or used (university/academic, clinical or both), the target population (e.g. pre-qualification, post-qualification or both), the format of the instrument (e.g. self-report questionnaire), an assessment of the instrument development process against the Standards and an indication of the construct measured as described by the typology of IPE outcomes described by Barr et al. Part B was used to evaluate the psychometric properties of the instrument. Part B facilitated the extraction of data on validity evidence and reliability, and information relating to score scales, score interpretations and score comparability. Further, an overall judgement of the extent to which the Standards had been met for each category was made. Data extraction used a spreadsheet developed on the basis of the QuAILS.

To ensure consistency of understanding of the QuAILS checklist, the authors independently extracted data for two included instruments and compared results. Discussion resulted in minor amendments to the QuAILS data extraction form. Data extraction was then performed by the primary review author and verified for each instrument by the secondary review author.

### RESULTS

The flow of papers through the review is illustrated in Fig. 1. A total of 127 instruments were identified by the CIHC and Thannhauser et al. The search of databases identified 548 abstracts, from which 35 full-text articles were retrieved. From this, 13 additional instruments were identified. The final instrument pool consisted of 140 instruments, nine of which met the inclusion criteria for critical appraisal.

### Qualitative attributes of instruments

The qualitative characteristics of each instrument are presented in Table 2. All instruments were self-report questionnaires, except the Interprofessional Collaborator Assessment Rubric (ICAR), which uses an observational format. Two instruments measured IPE outcomes at Level 1 (learner reactions),
eight at Level 2a (modification of attitudes), five at Level 2b (acquisition of knowledge/skills) and two at Level 3 (behaviour change) of Barr et al.’s framework of IPE outcomes.

The Standards relating to instrument development require developers to provide users with sufficient information relating to the procedures used to develop, review and trial the instrument. The processes by which items have been selected from an item pool and the model (classical test theory or item response theory) used for psychometric evaluation of the instrument are also set as standards for instrument development. The Interprofessional Socialisation and Valuing Scale (ISVS) was deemed to meet the standards relating to instrument development. Five instruments partially met these standards (ICAR, KidSIM ATTITUDES Questionnaire, Generic Role Perception Questionnaire [GRPO], the University of the West of England Interprofessional Questionnaire [UWEIPQ] and three did not meet these standards (Attitudes to Shared Learning [ASL], RIP-LS, StudData Questionnaire).

Figure 1 Flow of papers on instruments to measure outcomes of interprofessional education through the review.

Appraisal of the psychometric properties of instruments

A summary of the evidence for validity, reliability, scaling and scoring is presented in Table 3.

Validity evidence

Arguably the most important characteristic of any test, instrument or scale, validity is the degree to which available evidence serves to support the intended interpretations of instrument or test scores. The Standards identify five sources of validity evidence: test content; response processes; internal structure; relations to other variables, and consequences of testing.

Validity evidence based on test content

This evidence seeks to support the relationship between the instrument’s content (its items and their wording and format) and the construct it is intended to measure. Sources of this evidence include the views, ideas or opinions of experts in the field or of those who represent the population.
<table>
<thead>
<tr>
<th>Purpose</th>
<th>Instrument structure</th>
<th>Standards for instrument development</th>
<th>IPE outcomes*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interdisciplinary Education Perception Scale (IEPS) (Leucht et al.)</strong>&lt;sup&gt;1&lt;/sup&gt;</td>
<td>To measure four attitudes important to interdisciplinary settings: 18 items; 4 attitude subscales: Professional competency and autonomy (8 items); Perceived need for professional cooperation (2 items); Perception of actual cooperation and resource sharing within and across professions (5 items); Understanding the value and contributions of other professionals/professions (3 items)</td>
<td>Partially met 2a</td>
<td></td>
</tr>
<tr>
<td><strong>Readiness for Interprofessional Learning Scale (RIPLS) (Parsell &amp; Bligh)</strong>&lt;sup&gt;2&lt;/sup&gt;</td>
<td>To assess the ‘readiness’ of students for shared learning: 19 items; 3 subscales: Teamwork and collaboration (9 items clustered into 2 groups: Effective teamworking [6 items], Relationships with other professionals [3 items]); Professional identity (7 items: Negative professional identity [3 items], Positive professional identity [4 items]); Roles and responsibilities (3 items)</td>
<td>Not met 2a</td>
<td></td>
</tr>
<tr>
<td><strong>Attitudes to Shared Learning (ASL) (Forman &amp; Nyatanga)</strong>&lt;sup&gt;3&lt;/sup&gt;</td>
<td>To measure student attitudes to shared learning: 60 Likert-type items, 12 open-ended questions relating to experience. Items cover concepts, curriculum issues, statements about interprofessional/shared learning, social aspects of the course, curriculum aspects, problem-based learning, working practice, other professionals’ roles, support from institution, logistical aspects</td>
<td>Not met 1, 2a</td>
<td></td>
</tr>
<tr>
<td><strong>University of the West of England Interprofessional Questionnaire (UWEIPQ) (Pollard et al.)</strong>&lt;sup&gt;4, 5&lt;/sup&gt;</td>
<td>To measure student self-assessment of communication and teamwork skills, attitudes towards interprofessional learning, students’ perceptions of interaction between health professionals, students’ perceptions of their relationships with colleagues from their own and other professions: 35 items; 4 subscales: Communication and teamwork (9 items); Interprofessional learning (9 items); Interprofessional interaction (9 items); Interprofessional relationships (8 items)</td>
<td>Partially met 2a, 2b</td>
<td></td>
</tr>
<tr>
<td><strong>Generic Role Perception Questionnaire (GRPQ) (Makay)</strong>&lt;sup&gt;6&lt;/sup&gt;</td>
<td>To measure the perception of the role of a range of professions: 20 bipolar role construct items</td>
<td>Partially met 2a</td>
<td></td>
</tr>
<tr>
<td><strong>Interprofessional Socialisation and Valuing Scale (ISVS) (King et al.)</strong>&lt;sup&gt;7&lt;/sup&gt;</td>
<td>To measure aspects of the interprofessional socialisation process: 24 items; 3 subscales: Self-perceived ability to work with others (9 items); Value in working with others (9 items); Comfort in working with others (6 items)</td>
<td>standards met 2a, 2b, 3</td>
<td></td>
</tr>
</tbody>
</table>
Outcome measurement instruments in IPE

Table 2  (Continued)

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Instrument structure</th>
<th>Standards for instrument development</th>
<th>IPE outcomes*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>StudData Questionnaire (Almás &amp; Barr(^2), Almás &amp; Ødegård(^3))</strong></td>
<td>To measure central aspects of interprofessionalism as a construct: 16 items; 3 subscales: Need for interprofessional collaboration (6 items); Value of interprofessional education (7 items); Openness to interprofessionalism (3 items)</td>
<td>Not met</td>
<td>2a, 2b</td>
</tr>
<tr>
<td><strong>Interprofessional Collaborator Assessment Rubric (ICAR) (Curran et al.(^13))</strong></td>
<td>To assess (formatively and summatively) a learner’s achievement of stated interprofessional collaborator competencies: 6 competencies each with a number of dimensions (dims) and behavioural indicators (Bls): Communication (2 dims, 7 Bls); Collaboration (3 dims, 4 Bls); Roles and responsibilities (4 dims, 7 Bls); Collaborative patient/client family-centred approach (4 dims, 4Bls); Team functioning (3 dims, 5 Bls); Conflict management/resolution (3 dims, 4 Bls)</td>
<td>Partially met</td>
<td>2b, 3</td>
</tr>
<tr>
<td><strong>KidSIM Attitudes Towards Teamwork in Training Undergoing Designed Educational Simulation (ATTITUDES) (Sigalet et al.(^15))</strong></td>
<td>To measure student perceptions of and attitudes toward IPE, teamwork, and simulation as a learning modality: 30 items; 5 subscales: Communication (8 items); Relevance of IPE (7 items); Relevance of simulation (5 items); Roles and responsibilities (6 items); Situation awareness (4 items)</td>
<td>Partially met</td>
<td>1, 2a, 2b</td>
</tr>
</tbody>
</table>

* Interprofessional education outcomes: Level 1 (learner reactions); Level 2a (attitudes/perceptions); Level 2b (knowledge/skills); Level 3 (behaviour); Level 4a (organisational practice); Level 4b (benefits to patients) (based on Barr et al.\(^8\)).
† Revised version of IEPS developed by McFadyen et al.\(^25\).
‡ Revised versions: McFadyen et al.\(^26\), Lauffs et al.\(^27\) (Swedish translation); Wilhelmsson et al.\(^29\) (modified Swedish translation); Williams et al.\(^20\) (Rasched); Tamura et al.\(^28\) (Japanese translation).

in which the instrument is intended to be used (e.g. health professional students) or of people who receive interprofessional care.

Information relating to the development of content was reported for all instruments. Instrument development typically involved the use of literature reviews and experts in the field of IPE to conceptualise the domains of IPE to be measured, and to generate a pool of test items. No information was reported outlining the criteria for quality item design. Instruments were field-tested and trialled using pre-qualification students from a range of health professional disciplines. With the exception of the ICAR, GRPQ and IEPS, students did not have any direct input to the development of instruments and no instrument included patient input.

**Validity evidence based on response processes**

This type of evidence is concerned with the relationship between the construct or outcome being measured and the response of those using the instrument. This evidence is generally sourced from analyses of users’ responses to individual items or the whole instrument through ‘think aloud’ or exit interviews.\(^24\) Such evidence is particularly relevant when observers are used to judge performance against assessment criteria, as is often required to determine the competency of health professional students in their profession. Evidence that such judgements are being made without influence by factors irrelevant to the interpretation of the criteria is important. No reports of this type of validity evidence were found.

**Validity evidence based on internal structure**

Evidence based on internal structure provides an indication of the interrelationship of items in a test or instrument. This enables instrument designers to understand the degree to which all the items of an instrument relate to the intended construct. Factor analysis was used to identify subscales in six instruments (ISVS, KidSIM, IEPS, RIPLS, UWEIPQ, and StudData).
StudDATA). Subsequent factor analysis of the IEPS by McFadyen et al.\textsuperscript{25} removed the fourth factor and six items from the original IEPS developed by Leucht et al.\textsuperscript{7} Similarly, McFadyen et al.\textsuperscript{26} proposed a revised four-subscale version of the RIPLS following factor analysis. The Swedish version of the RIPLS developed by Lauffs et al.\textsuperscript{27} retained the original three-factor model, as did the Japanese version developed by Tamura et al.\textsuperscript{28} However, factor analysis of the Swedish version of the RIPLS developed by Lauffs et al.\textsuperscript{27} by Wilhelmsson et al.\textsuperscript{29} proposed a new four-factor structure. In their Rasch analysis of the RIPLS based on a four-factor model, Williams et al.\textsuperscript{30} removed two items (items 4 and 9) found not to fit to the Rasch measurement model (RMM).

Five instruments (ASL, ISVS, KidSIM, UWEIPQ, StudDATA) reported Cronbach’s alpha values for each subscale or factor within the acceptable range of 0.70–0.90 suggested by Streiner.\textsuperscript{31} The remaining four instruments included subscales or factors for which Cronbach’s alpha values lay out with the acceptable range (RIPLS, IEPS), or were not reported (GRPQ), or were measured inappropriately (ICAR).

Validity evidence based on relationships to other variables

This source of evidence is concerned with the degree to which an instrument’s scores are correlated to scores obtained by other, concurrently administered measures of the same or similar construct (convergent validity) and not correlated with dissimilar or unrelated variables (discriminant validity). If, for example, the scores obtained by an instrument which measures outcomes of IPE in pre-qualification health professionals were found to predict interprofessional collaborative practice performance as directly observed and assessed by an assessor, this would contribute to the validity evidence of the instrument.

This type of evidence was found for only the UWEIPQ. A correlational study examined the concurrent validity of subscale 1, the Communication and Teamwork Scale (against the Interpersonal Communication Competence Scale\textsuperscript{32}), subscale 2, the Interprofessional Learning Scale (against the RIPLS), and subscale 4, the Interprofessional Relationships Scale (against the IEPS) and found Pearson correlation coefficients of 0.85, 0.84 and 0.72, respectively.\textsuperscript{17} However, whether the participant scores on the RIPLS and IEPS used for these correlational studies were raw scores, subscale scores or total scores is unknown.

Validity evidence based on consequences of testing

The Standards propose that the intended and unintended consequences of an instrument’s use be examined. For example, in the context of an instrument designed to measure outcomes of IPE, where such measurement is anticipated to lead to improvements in course or curriculum design, evidence that these benefits are being realised would be included in the process of validating the instrument. Where such instruments may be used to provide formative feedback to students about their learning, the consequences of such use would need to be examined. None of the instruments reviewed provided evidence based on consequences of testing.

Reliability

Reliability is concerned with both the dependability of test scores, inferred by the consistency of test scores over repeated administrations of a test, and the degree to which scores are free of errors of measurement.\textsuperscript{11} The score obtained by a scale or instrument (observed score) may differ from the real score (true score) for a respondent. This true score may not be observed when error, relative or absolute, is associated with the measurement. Relative error is of particular relevance given the self-report nature of instruments used to measure outcomes of IPE.\textsuperscript{33} When a self-report instrument or scale is administered to a cohort of participants, any differences in individual scores result from relative differences between individuals within the participant group rather than variation in the administration of the instrument or scale. Absolute error would be of interest where the administration of an instrument is subject to variation. An example of this would be the observed judgement of skill or performance of a task using different assessors across a participant group. If we are to be able to evaluate an instrument and have confidence in its use, information about potential sources of measurement error is important.

Reliability information may be reported as standard deviations of measurement errors, the standard error of measurement (SEM), or in the form of one or more reliability coefficients. Only two instruments, the GRPQ and the UWEIPQ, provided relative reliability data in the form of correlation coefficients. No reporting of SEM data was identified.
Table 3  Summary of evidence of validity, reliability, scales and scores in instruments to measure outcomes of interprofessional education

<table>
<thead>
<tr>
<th>Evidence</th>
<th>ASL</th>
<th>ISVS</th>
<th>ICAR</th>
<th>ATTITUDES</th>
<th>GRPQ</th>
<th>IEPS</th>
<th>RIPLS</th>
<th>UWE</th>
<th>IPQ</th>
<th>StudDATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) Test content</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was a pool of items generated?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Were criteria for item pool reduction specified?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Were criteria for item design quality specified?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Was the instrument trialled in an environment aligned with the intended purpose of the instrument?</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Were characteristics of study participants reported?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Were characteristics of participants involved in the development of the instrument reported?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Were pre-qualification health students involved in the development of the instrument?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2) Internal structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have statistical approaches been used to investigate the internal structure of the instrument?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Was Cronbach's alpha reported for each subscale or factor and within acceptable range 0.70–0.90*?</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Was any bias for items among different subgroups investigated?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3) Response processes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have analyses of individual responses been completed?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>4) Relationship to other variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have any correlational studies been completed?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>5) Consequences of testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have the benefits of using the instrument been described?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Overall assessment against Standards relating to validity</td>
<td>–</td>
<td>/</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has reliability been measured and reported?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Have measurement error data been reported?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Overall assessment against Standards relating to reliability:</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>/</td>
<td>–</td>
<td>/</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Scales and scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has information relating to the construction of score scales been provided?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>
Scales and scores

Instrument developers should report information in relation to the construction of score scales, the protocol for calculating scores and a description of the interpretation and meaning of scores for potential users. Only the ISVS and UWEIPQ satisfied these obligations. Although information relating to the construction of score scales was generally reported, the score calculation protocol and the interpretation and meaning of particular scores were less frequently specified. The IEPS and RIPLS, being the most widely adopted instruments, demonstrated inconsistency in the application of their respective scoring protocols among test users.

DISCUSSION

The purpose of this review was to critically appraise the qualitative and psychometric properties of instruments that measure outcomes of IPE in pre-qualification health professionals. This review has failed to identify an instrument or instruments that meet the criteria set by the Standards for Educational and Psychological Testing for the design and development of educational tests.

Validity evidence

The Standards require that appropriate evidence of validity be gathered to mount an argument to support the intended interpretations of instrument scores. Instrument users must be confident in the internal structure of a scale or subscale, particularly where items are summed to provide a total score. Where a total score is indicated, users must be satisfied that the scale is unidimensional. Evidence based on internal structure is reported in terms of factor analysis data and internal consistency data, in the form of Cronbach’s alpha or on the basis of data fit to a model in item response theory.

Factor analysis is used to explore the structure of an instrument. With the exception of the IEPS and RIPLS, the subscale structure has remained stable over time in all instruments in this appraisal. Further factor analysis of the IEPS reduced the original four subscales to three subscales, and the work of various authors either retained the original three-subscale version of the RIPLS or proposed a revised four-subscale version. Cronbach’s alpha is a useful measure of an instrument’s internal consistency, but it must be used and interpreted with caution. Acceptable Cronbach’s alpha values may not reflect the internal consistency of items or unidimensionality of the scale, but may derive from a large number of test items.

Developers of the UWEIPQ provided some evidence of validity in relation to other variables. Although these values suggest a strong relationship between the relevant subscale of the UWEIPQ and the comparator scale, some caution is necessary when con-
sidering the value of this evidence. Given that any comparison of the UWEIPQ subscales and the other scales involves the comparison of two self-report measures of a similar construct, the weight of this concurrent validity evidence is questionable. The strength of this evidence would be greatly improved by correlating the UWEIPQ with a different measure of the same construct, such as measures based on ratings by others or direct observation. 34

Instrument developers should consider other potential sources of evidence to support the validity of interpretations of scores. Based on their experience of interprofessional and collaborative health care, patients can provide rich and meaningful perspectives that can contribute to our understanding of the constructs relating to interprofessional collaborative practice and, in turn, outcomes of IPE. Similarly, the use of pre-qualification health professional students in the instrument design process may yield more content-valid scales. It is often assumed that scores on an instrument may predict future performance. Such analysis involving the assessment of students in clinical or practice-based contexts may be a useful adjunct to the body of validity evidence.

Reliability

Measures of reliability are important components of robust instruments and scales designed to measure educational outcomes, particularly when they are used to measure changes in IPE outcomes in pre/-post-intervention studies or to detect differences between groups in controlled studies. If scores collected on two occasions are not consistent (despite no change in the underlying construct of interest), then the ‘signal’ of differences or changes in the construct cannot be detected over the ‘noise’ of the score variability or measurement error. To date, little or no attention has been given to examining the reliability of the instruments reviewed.

Scales and scores

The IEPS and RIPLS are the instruments most frequently utilised by researchers in the field of IPE and collaborative practice. The IEPS and RIPLS have been used to evaluate the effectiveness of IPE interventions such as interprofessional curricula,29,35–43 IPE projects,44–48 interprofessional clinical or fieldwork placements,29,49–56 and online IPE learning modules.57,58 Although both instruments have been popular choices for researchers and educators alike, their use in measuring between-group differences or within-group change resulting from IPE interventions is problematic.

Problems relating to the scoring of test items are evident with both instruments. In the case of the IEPS, the original developers used a weighted scoring system to derive an individual respondent’s test score. However, subsequent use of the IEPS by various authors demonstrates considerable variation in the application of the original test scoring protocol. In some papers, raw scores have been used for analysis rather than factor or subscale scores and total scores. In the RIPLS four negatively worded items must be reverse-scored in order to determine a respondent’s score. This is not clear from the original publication reporting on the development of the RIPLS.6 Many subsequent users of the RIPLS have neglected to reverse-score responses to these negatively worded items.

The inconsistent and inaccurate application of item and test scoring by adopters of the IEPS and RIPLS makes it difficult to compare findings across studies that use these instruments. Modified versions of these instruments, with different subscales, add to the difficulty of drawing comparisons between published studies. Furthermore, such modified versions may necessitate the collection of additional validity evidence to support the intended interpretations of test scores obtained when using any modified versions of instruments.

Test construction paradigm

In studies that have used the IEPS and RIPLS to evaluate Level 2a outcomes (attitudes and perceptions) of IPE interventions, authors frequently report no significant difference between pre- and post-IPE intervention mean scores, which may reflect poor sensitivity of the IEPS and RIPLS to detect meaningful change over time.29,40,42,44,46,54,58 The use of classical test theory to construct all instruments in this review may provide some explanation for this finding. When developing a set of items that make up an instrument, a continuum on which respondent ability and item difficulty are located for a particular trait (e.g. attitudes to interprofessional learning) should be developed. Items must be able to discriminate between respondents with lower (e.g. ambivalent or less positive attitude) and greater (e.g. highly positive attitude) ability. Classical test theory does not scale items in this way. Rather, items tend to be equally difficult and contribute equally to the respondent’s total test score. This can create a ceiling effect, which makes dis-
discriminating between those respondents in the upper range of the instrument’s scale difficult.

Alternative test construction paradigms have evolved in response to the recognition of the underlying weakness of classical test theory.\textsuperscript{59} Item response theory and the RMM, although developed independently and possessing different characteristics, address the weakness of the classical test theory model. Proponents of the RMM believe it has greater potential than item response theory in the field of health measurement and ultimately to improve patient care.\textsuperscript{50} The reasons for this are complex and beyond the scope of this article. However, more advanced measurement models are available and, if applied correctly, may yield more generalisable and discriminating scores.

None of the instruments included for review in this appraisal were developed using advanced measurement models. However, recent work by Williams \textit{et al.} used the RMM to investigate the psychometric properties of the original RIPLS.\textsuperscript{30} Using the RMM, Williams \textit{et al.}\textsuperscript{30} proposed a four-factor model (as opposed to the original three-factor model developed by Parsell and Bligh\textsuperscript{6}). Following the removal of two misfitting items in subscale 1 (shared learning), there was good fit to the RMM.

\textbf{Future directions}

Future work in scale development in IPE should investigate all sources of measurement error, rather than only those associated with the sampling of items. The use of alternative measurement models (e.g. the RMM) may yield more precise estimates of IPE traits of interest.

\textbf{Limitations}

This review considered only those instruments designed for the measurement and evaluation of IPE in pre-qualification contexts. The outcomes of this review may not be generalisable to other contexts, such as post-qualification IPE.

\textbf{CONCLUSIONS}

This review and critical appraisal suggest that there is insufficient psychometric evidence (on validity and reliability) to support the use of currently available instruments to measure outcomes of IPE in pre-qualification health professional students. Further scale development and validation work is necessary.


**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article:

**Data S1.** Quality Appraisal of Interprofessional Learning Scales (QuAILS).

**Appendix S1.** Description and user guide for the Quality Appraisal of Interprofessional Learning Scales (QuAILS), a standardised checklist specifically developed for the purpose of this review.

Received 8 July 2014; editorial comments to author 8 September 2014, 28 November 2014; accepted for publication 15 December 2014.
Copyright of Medical Education is the property of Wiley-Blackwell and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.