Assessing the Assessment: using psychometrics to establish reliability and validity of portfolio assessment

Anders Beckman, Institutionen för kliniska vetenskaper I Malmö

Handledare: Ann-Christin Haffling
Assessing the Assessment: using psychometrics to establish reliability and validity of portfolio assessment

Introduction
Assessment
Assessment is an essential component of education, with diverse aims (1). In medical education assessment is of interest for students (e.g. motivation and progress), teachers (e.g. control of teaching) and patients (e.g. professionalism of graduates) and has moved from assessment of learning to assessment for learning (2). The usefulness is determined by the assessments’ standard in fulfilling good criteria, i.e. to be valid, reliable, feasible, fair and beneficial to learning (3).

Portfolio assessment
Several instruments are regarded as good assessment tools, due to a substantial body of evidence (4). However, attributes like attitudes, values, aptitude for communication, clinical reasoning and self-reflection (5), are not easily assessed by traditional instruments. A possible tool for assessment of these dimensions of professional competence is the portfolio, provided it contains self-reflection and not only documentary evidence from clinical practice (6). The reflection should include analysis of educational achievements, identification of further learning needs, and a learning plan (7).

Portfolios should be assessed summatively to compete on equal conditions with other assessment instruments (8). However, reliability and validity of portfolios have been questioned, and hence the use of portfolios for summative assessment (9) Customary psychometric measures of reliability and validity for portfolios in medical education are not uniformly used and the interpretation is diverse.

Moreover, due to the fact that assessment of a reflective portfolio is mainly an assessment of qualitative substance, alternative methods to conventional psychometrics have been proposed; such as using procedures, developed from qualitative research criteria (10). By this tactic, ‘dependability’ (reliability) can be used for the assessment process, and ‘credibility’ (internal validity) for the assessment structure.

Reliability
Reliability is the criteria equivalent to instrumental measurement precision, the consistency of repeated tests or the reproducibility and repeatability (2, 11). In portfolio assessment it can be enhanced by standardization of portfolio content, definition of analytical assessment criteria, clear guidelines to students (12) and discussion and negotiation between examiners (13). Reliability can also be improved by a personal examiner interview (12).

In the assessment of portfolios, the inter-rater-reliability is the main issue. The terminology and the tools for analysis of inter-rater reliability are not definite (14). The intersecting area in medical teaching of medical and social science offers a field for discussion. Clear-cut methods and procedures must be established, for the portfolio to be regarded as a good summative assessment tool.

Definitions and methods
Recently guidelines for inter-rater reliability and agreement have been proposed (15), albeit in the field of medical science. However, these guidelines can be used to improve reporting in the field of medical education.
There is a distinction between inter-rater agreement (IRA) and inter-rater reliability (IRR) (16) (17). IRA is the extent to which assessors make exactly the same judgment about a subject (18), i.e. the scores of different assessors are interchangeable. IRA equals the terms ‘consensus estimate’ (19) or ‘absolute consensus’ (20).

IRR is concerned with the ability of scores to distinguish between subjects. The IRR expresses how well different assessors’ ratings correlate; if there is a consistency in the rating of subjects, although each subject is not provided exactly the same rating by different assessors (19, 20).

IRA can be estimated by various methods (19) (21):

- **Percent agreement** is implemented by adding cases with the exact same rating/score and dividing by the total number of cases. It is easy to do by hand but it does not take account of differences due to chance.
- **Cohen’s kappa** is an index that estimates degree of consensus between two judges, originally developed for nominal data but not correcting degree of discrepancy. To allow for adjustment of discrepancy, weighted kappa can be used (22). An analysis for more than two assessors can be made by Fleiss’ kappa (23, 24). Values greater than 0.75 are regarded as excellent agreement (25).
- **Intraclass correlation coefficient (ICC)** is used to estimate error/reliability with numerical variables (25). It measures the ratio between observed and “true” variance. However, the calculation of ICC presumes an interval scale. Furthermore, ICC contains both intra- and inter-rater reliability and is therefore hard to interpret.

IRR can be valued by other methods:

- **Pearson’s** correlation coefficient is a coefficient between -1 and +1, requires a normally distributed interval scale, and can only be calculated for one pair of assessors at a time.
- **Spearman’s** ranking coefficient is the equivalent to Pearson’s for not normally distributed, ordinal data. Again, calculation is only possible for two assessors at a time.
- **Cronbach’s alpha** is a form of correlation coefficient that measures the variance in a scale or of assessors. It is widely used for the testing of internal consistency when several assessors rate the same students (26). Prerequisite for this coefficient of homogeneity is that the area of interest has a common construct (27).
- Some sources say that ICC can be used for IRR, depending on the model used (28, 29)

A third type of estimation is suggested by Stemler (19): The measurement estimate. This concerns the ability to create a summary score of different assessors, particularly when the rating scale represents different levels of competence and/or not all assessors rate all items. Main methods are factor analysis and Rasch model (19).

**Inter-rater reliability of the portfolio**

Several methods have previously been used in different settings to assess the inter-rater reliability of portfolios; the use of Fleiss’ kappa with a dichotomous scale for GPs in training resulted in only a “fair” (k=0,1-0,41) inter-rater reliability (30-32); the use of intraclass correlation between two assessors on a summed and percentage converted score resulted in a high correlation (ICC=0,771) between raters (13); the use of Spearman’s Rho for comparing the ratings of students by pairs of assessors resulted in highly varying levels (0,16-0,65) (33).
Regarding the testing of inter-rater reliability, these examples show methodological problems, not fully emphasized: the number of students must be sufficient (enough “power”); test-methods must be able to incorporate more than two assessors; and depending on the assessment scale, the method must be appropriate.

Validity
Validity of an assessment refers to “the extent to which an assessment measures what it is supposed to measure” (1). Validity has several different aspects, which can have varying degrees of fulfilment (1).

Definitions and methods

- **Content validity** refers to the degree to which the instrument measures students’ learning outcomes. The content validity can be judged by professionals and can be increased if relevant tasks are measured, if appropriate formats are selected, and if the test is scrutinized by reviewers (3).

- **Criterion validity** is a term used for concurrent as well as and predictive validity (11). *Concurrent validity* denotes how well the results from an assessment instrument correlates with an established instrument. This can be judged by correlating scores with another course test by using standard instruments such as Pearson’s or Spearman’s Rho (3). *Predictive validity* refers to if the instrument can predict performance in the future and can warrant long time to valuate. It is not a widely used validation (2).

- **Construct validity** regards the assessment of traits that are hypothetical or not observable (e.g. empathy, creativity, reasoning) (34) (35). The construct validity can be *convergent* if it correlates well with a test that assesses the same construct, and it can be *divergent* if it has a low correlation with a test that assesses another construct, thereby implying good validity in spite of low correlation! One way to test construct validity is to use factor analysis. (11) Factor analysis shows variability between different variables, and analyses if there is a common factor in an exam, between items or between exams. It is a form of a data reduction to reveal unobserved (latent) variables, i.e. if a single factor is assessed with different assessments or if different factors are assessed (21).

- **Face validity** involves the appearance of the assessment to examiners and teachers. It can be corroborated by student and teacher feedback (1).

Validity of the portfolio
Earlier work regarding portfolios in medical education is sparse in the direct reporting or discussion of validity. Consequent and comprehensive reporting and/or discussion of the different aspects of validity are hard to find.

In a composite assessment process for internal medicine clerkship in Cape Town, South Africa (31), medical students were assessed using a portfolio, an in-course assessment, a MCQ-test, and a clinical examination. The Pearson’s correlation of students’ portfolio scores with scores of other components of the assessment where low to moderate, but the authors only discussed this in the area of concurrent validity. Construct validity was not mentioned (36).

In a final examination for medical students in Dundee, Scotland, one part consisted of a portfolio assessment. Validity was only discussed in the realm of content validity and a hint of divergent construct validity. Spearman’s correlation coefficient was calculated between students’ scores of the portfolio and other tests. Surprisingly, the authors discussed the results of the correlation in terms of reliability, not concurrent validity (14).
Aim
In summary, previous studies of portfolios in medical education have tested inter-rater reliability by the use of varying methods without problematizing the difference between IRA and IRR, thereby neglecting the choice of the best method. Furthermore, reliability has predominantly been tested with two assessors, a setting not fit for medical education with a large number of students and assessors. Although validity by some authors is regarded as a dynamic process and specific for a certain process (2), different aspects of validity of the portfolio deserves a better discussion.

In a recently published paper on a portfolio pilot, undergraduate students’ reflections on professional issues in general practice were qualitatively analysed (37). The authors were able to show sufficient results of reliability and internal validity of the portfolio by the use of criteria for qualitative research, and content validity was also supported by the content of students’ reflections. However, the inter-rater reliability, the concurrent validity, and possible effects of the self-selected sample of students warranted further psychometric evaluation. This study aims to analyse these criteria, using different traditional psychometric methods.

Method
A full outline of the portfolio has previously been reported, and below only a short summary is provided (37).

Participants
Medical students of two consecutive classes of the final year at Lund University were offered to voluntarily pilot a portfolio in general practice as an addition to the written test. A total of 35 students accepted, 24 (69%) of whom were women. The students practice 16 days in a health centre, and the practice involves a high degree of authentic learning in the workplace. The portfolio thus appeared to be an instrument, well suited for formative as well as summative assessment of professional development. Four course teachers, of whom two were women, acted as mentors for their students and later as examiners for each others’ students.

Portfolio
Learning outcomes and content of the portfolio are provided in Box 1. The content was standardised to include documentary evidence from practice; including self- and teacher formative assessment, eight-item checklist of practical skills, documented feedback from video seminars of recorded video consultations, and case summaries of two patient encounters, illustrating characteristics of general practice. Based upon the documentation, students also wrote three reflections; one reflection on their consultation skills and two reflections on their two case summaries.
Learning outcomes
The student is expected to
- analyse his/her strengths and weaknesses in general practice consultations, reflect on learning achieved, what areas need further development and how this development may take place
- discuss and critically examine characteristic attributes of general practice, illustrated by two case summaries from practice, reflect on learning achieved, and on how further development may take place

Content

Documentary evidence
- Assessment protocols (self-assessment and feedback) from practice
- Checklist of practical skills from practice
- Feedback sheet from video seminar
- Two case summaries, critical incidents or routine clinical experiences, illustrating characteristic attributes of general practice

Reflections
- One reflection on consultation skills, based upon documentation from practice and video group
- Two reflections, based on the two case summaries, and linked to characteristic attributes of general practice

Box 1 The portfolio: learning outcomes and content

Rating process and criteria
Except for the case summaries, the documentary evidence of the portfolio was formatively assessed in practice or in seminars, and the summative assessment related to the three reflections. Assessment in medical school, Lund University only includes pass/fail. The examiners read students’ portfolios and prepared for the personal 30 min interview that completed the assessment. The interview had a holistic approach; students were challenged to develop on their reflections, and received feedback.

To estimate the inter-rater reliability of the assessment, all four teachers three months after the last exam separately read and rated all the 35 portfolios. The assessment concerned two main areas: the three reflections and the portfolio as a whole (Fig 1).

For each of the three reflections ‘learning’ and ‘links to documentary evidence’/‘general practice attributes’ was evaluated, using a 4-point ordinal scale (1=poor, 4=excellent); and structure and written language was assessed for the portfolio as a whole, implying theoretically a minimum of 7 and a maximum of 28 points of each portfolio. The rating criteria are listed in Box 2.

Written exam
All portfolio students participated in the written exam, and their results were for analytical purposes normalised to a percentage scale.
<table>
<thead>
<tr>
<th>Summative assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reflections</strong></td>
</tr>
<tr>
<td>-Consultation</td>
</tr>
<tr>
<td>-Case summary 1</td>
</tr>
<tr>
<td>-Case summary 2</td>
</tr>
<tr>
<td><strong>Learning</strong></td>
</tr>
<tr>
<td>-Use of documentary evidence (general practice attributes)</td>
</tr>
<tr>
<td><strong>Whole portfolio</strong></td>
</tr>
<tr>
<td>-Structure and written language</td>
</tr>
</tbody>
</table>

Figure 1. Summative assessment
Consultation skills - Learning
1. No reflection on learning. Students discuss consultation skills, but do not reflect on their learning.
2. Limited reflection on learning. Students describe their strengths, but critical reflection on what needs to be improved is missing.
3. Fairly good reflection on learning. Students discuss their development, but only some of students’ weaknesses and areas that need to be improved are mentioned.
4. Excellent reflection on learning. Students discuss their development, strengths and weaknesses are commented upon, and the reflection is applied to future experience.

Consultation skills - Links to documentary evidence
1. No links to documentary evidence.
2. Limited links to documentary evidence. There is an attempt to make a few links, but the result is unclear.
3. Fairly good links to documentary evidence. A fair range of documentary evidence is cited, but some important parts are omitted or some links are unclear.
4. Excellent links to documentary evidence. An extensive range of documentary evidence is cited and used in depth in the reflection.

Case summaries (two reflections) - Learning
1. No reflection on learning. Students only state their experience.
2. Limited reflection on learning. Students describe their experience, what they thought they did well and did not do well, but there is no reflection on how to develop further.
3. Fairly good reflection on learning. Students discuss learning from the experience and how to develop further, but the reflection is not applied to future experiences.
4. Excellent reflection on learning. Students discuss learning from the experience, how to develop further, and the reflection is applied to future experiences.

Case summaries (two reflections) - Links to general practice attributes
1. No links. The patient is described without any links to general practice.
2. Limited links. There is an attempt to make a few links, but the result is an unclear connection to general practice.
3. Fairly good links. One or two attributes are discussed, but the connection is not completely clear. Some obvious links are not discussed.
4. Excellent links. Clear links to general practice attributes are discussed and used in depth.

Structure and written language (the whole portfolio)
1. The structure is unclear and incoherent; language is poor with several mistakes in grammar and spelling.
2. The structure is unclear in places; language is poor in places with a few mistakes in grammar and spelling.
3. The structure is mainly logical and coherent and language is fairly good with no disturbing mistakes in grammar or spelling.
4. The structure is outstandingly clear, logical and coherent and language is excellent.

Box 2. Rating criteria for the portfolio reflections
Data analysis
Data from the four assessors’ ratings of students’ reflections and of students’ scores of the written exam were analysed using SPSS 19.0 (SPSS Inc, software location Lund University) and an online calculator for percent agreement and Fleiss’ kappa (38).

Reliability
The inter-rater agreement (IRA) was analysed using percent agreement, a modified Fleiss’ kappa, and intra-class correlation coefficient (ICC).
To establish the level of inter-rater reliability (IRR) for students’ total portfolio scores, an attempt was made with ICC, but Cronbach’s α coefficient was the main evaluation of ‘consistency estimates’.

Validity
The correlation between students’ total scores of the portfolio and of their written exam were analysed using Pearson’s correlation coefficient and Spearman’s Rho to estimate concurrent validity.
A factor analysis of the written exam and the portfolio was undertaken to determine if a single factor or if different factors were assessed by the two exams, i.e. estimating construct validity (19).
To determine if the portfolio students differed from the rest of students (self-selection bias), mean test scores of the written test for portfolio students and for the non-participating students were calculated.
Results
All analysis was made with the total scores of students’ portfolios.

Descriptive
The mean of the four assessors’ total scores for each student was 20.8 points (range 11.3-27.5) in the portfolio exam with a nearly normal distribution (Fig 2).

The four assessors’ means were 19.2, 20.7, 21.0 and 22.1 respectively, and their range of scores was 10-28 (Fig 3).
Reliability

IRA
Percent of overall agreement: 0.11
Free-marginal kappa: 0.062
ICC (two-way mixed model; absolute agreement): 0.870 (95% CI 0.772-0.930) average measure; 0.626 (95% CI 0.458-0.769) Single measure

IRR
Pearson’s correlation coefficient varied between 0.528 and 0.874 (Table 1)

<table>
<thead>
<tr>
<th></th>
<th>Ass.1</th>
<th>Ass.2</th>
<th>Ass.3</th>
<th>Ass.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ass.1</td>
<td>1</td>
<td>.528</td>
<td>.617*</td>
<td>.557*</td>
</tr>
<tr>
<td>Ass.2</td>
<td>.528</td>
<td>1</td>
<td>.803</td>
<td>.778</td>
</tr>
<tr>
<td>Ass.3</td>
<td>.617*</td>
<td>.803</td>
<td>1</td>
<td>.874</td>
</tr>
<tr>
<td>Ass.4</td>
<td>.557*</td>
<td>.778</td>
<td>.874</td>
<td>1</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Spearman’s rank coefficient varied between 0.466 and 0.870 (Table 2)

<table>
<thead>
<tr>
<th></th>
<th>Ass.1</th>
<th>Ass.2</th>
<th>Ass.3</th>
<th>Ass.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ass.1</td>
<td>1</td>
<td>.466</td>
<td>.532*</td>
<td>.510*</td>
</tr>
<tr>
<td>Ass.2</td>
<td>.466</td>
<td>1</td>
<td>.785</td>
<td>.764</td>
</tr>
<tr>
<td>Ass.3</td>
<td>.532*</td>
<td>.785</td>
<td>1</td>
<td>.870</td>
</tr>
<tr>
<td>Ass.4</td>
<td>.510*</td>
<td>.764</td>
<td>.870</td>
<td>1</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

ICC (two-way mixed model; consistency): 0.890 (95% CI 0.816-0.940) average measure; 0.670 (95% CI 0.526-0.795) Single measure

Cronbach’s α: 0.89

Validity

The concurrent validity, estimated by correlation between each student’s total portfolio scores and their written exam scores was calculated to 0.238 (p=0.168), by Pearson’s correlation coefficient and 0.308 (p=0.072) by Spearman’s rho.

The construct validity was estimated with factor analysis (principal component analysis with rotation varimax) of students’ portfolio scores and their written exam scores, and showed two components, with eigenvalue of 1.24 and 0.76 respectively, and percent of variance 62 and 38 respectively, illustrated in a component plot in rotated space (Fig. 4).
Figure 4 Component plot in rotated space of factor analysis

**Test for self selection bias**
The mean score of the written exam was 66.5 (95% CI 64.5-68.4) for portfolio students and 64.6 (95% CI 63.2-66.0) for the rest (n=105), i.e. no significant difference (p=0.17) (Fig. 5).

Figure 5 Mean score on written exam for portfolio students (P) and non-portfolio students (\,00)
Discussion

In this elaboration of a previously reported portfolio pilot, traditional psychometrics was used to illustrate different aspects of reliability and validity. The analysis showed high inter-rater reliability (consistency) by Cronbach’s alpha, low concurrent validity, compared to the results of the written test, evidence for high construct validity by factor analysis, and finally no sign of biases of the self-selected sample of students.

Inter-rater agreement (IRA), i.e. absolute agreement, is beneficial when clear-cut grades are important or when nominal categories are dealt with. Two tests of our own portfolio pilot for IRA, percent agreement and (modified) Fleiss’ kappa, showed very poor absolute agreement. This was expected, as percent agreement and kappa analysis of Cohen and Fleiss are primarily intended for nominal, categorical data, not for an ordinal scale (23). However, ICC as a third measure of IRA showed a very high value, a questionable result, illustrating that the (summative) ordinal scale had a great resemblance to normally distributed data, mocking an absolute agreement. Analysis of IRA with ICC can be done in many different ways (40) and hence the results can vary. Moreover, depending on the conditions, the ICC is a composite measure of both intra- and inter-rater reliability. Here is obviously a pitfall in reporting reliability, although other authors suggest ICC as a good substitute for the more complicated generalisability theory (GT) (41).

A high IRA is essential if the assessment is used for pass/fail, illustrated by the work of Pitts et al. (30-32). The analysis of inter-rater agreement for the dichotomous grade was in these three papers made by Fleiss’ kappa, but results were poor to fair. However, validity of the portfolio in question was only briefly discussed, and instructions to the participants appear to have been sparse, which was reflected by the abundant variation of size and content of their portfolios. In a summative assessment it is fundamental that clear learning outcomes are assessed (27). The reliability is increased if instructions are explicit, and if the structure of the portfolio is standardized (12).

A more elaborate study was made by Rees & Sheard in Nottingham, UK (13), where a substantial number of structured portfolios (one hundred) were assessed by two assessors, using explicit criteria. However, an ordinal scale was used for assessment, and a continuous scale was used for analysis by ICC. This transformation of an ordinal scale to a continuous scale and hence the use of parametric instead of non-parametric methods could be questioned, although this approach is sometimes recommended (39).

In our portfolio assessment of students’ reflections on professional practice, the inter-rater reliability as a consistency estimate (IRR) was our main interest; i.e. the aptitude of the assessors to equally, but not with absolute agreement, evaluate the students. One of the largest menaces to the continuing use of summative portfolio assessment is a low inter-rater reproducibility. In a fairly large study of portfolio assessment, Grant et al. found varying levels of inter-rater reliability, depending on pairs of assessors compared (33). The reliability was analysed using Spearman’s rank coefficient, a suitable method for the ordinal scale used. However, since Spearman’s rank coefficient can only be used for pairs of assessors, the overall reliability could not be evaluated. On the other hand it was easy to identify the poorest performing assessors.

In our study, the mediocre to good results of Pearson’s and Spearman’s correlation coefficients confirm the results of Grant et al. (33) in the downside of testing only one pair of assessors at a time. Depending on the choice of pairs, correlation can be high or low. By allowing for all assessors’ grading to be tested simultaneously by the use of Cronbach’s alpha, we were able to reduce the selection bias. The internal consistency reliability showed a remarkably high value. There is a risk of inflation of Cronbach’s alpha when only few grades are used in assessment, especially if the sample size is low (42). However, this risk of inflation is described for much greater item reduction than was the case in our study.
Furthermore, Cronbach’s alpha is easy to compute with modern software programs, and concern has been voiced about interpretation as a complex issue, that should warrant more elaboration (43).

Except for cognitive, technical, relational and integrative dimensions, professional competence also includes an affective dimension, ‘professionalism’; a complex phenomenon, including values, attitudes and behaviours (44). Portfolios are supposed to have the potential to evaluate students’ professionalism, as portfolios emphasise reflective practice (45). It is obvious that students’ results of a portfolio for reflections on professional practice indicate something else than their written exam that assesses cognitive knowledge. In our pilot, the low correlation between students’ results of the portfolio and the written exam was demonstrated by parametric and non-parametric methods, and the concurrent validity was low. However, we were able to establish evidence of construct validity by the result of factor analysis, confirming that the two exams assessed different components or constructs.

Although several instruments for assessment of learning outcomes have evolved, the instruments at best only reaches as far as estimating the ‘relational’ level of learning (46). For the level of ‘extended abstract’, i.e. students’ ability to analyse, assess, discuss and reflect, portfolios could fill a gap (46). In our setting, the written test with key feature problems assessed ‘relational’ learning outcomes and the portfolio assessed ‘extended abstract’ learning.

Strength of this study is the demonstration of different aspects of reliability and validity by the use of different psychometrics for analysis. Furthermore, the study also confirms the usefulness of the portfolio as a completing instrument for summative assessment of attributes, not possible to assess by other traditional tools.

Nevertheless, limitations exist. Most importantly, the relatively small group of participating students, who also were self-selected, could possibly bias the results. However, the students’ efforts of their portfolios and their scores of the written exam were normally distributed, suggesting that these students did not differ from their fellow comrades. This fact was further corroborated by the absence of differences in mean scores of the written exam between the portfolio students and their non-participating peers. The selection of assessors from the active and interested staff could be argued to bias the estimation of the IRR. However, the four assessors’ individual involvement in teaching, their experience and their knowledge of medical education varied considerably, thereby making them fairly representative of all teachers. The different results of Spearman’s correlation coefficient, measured between different pairs of assessors, might also reflect ambiguous assessment criteria, insufficiently agreed upon.

By disentangling different aspects of reliability and validity, and by using different psychometrics for analysis, this study has demonstrated that portfolio assessment for summative purposes is possible; however, appropriate methods must be used for an adequate evaluation.
Referenser

43. Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach’s $\alpha$, Revelle’s $\beta$, and Mcdonald’s $\omega$: their relations with each other and two alternative conceptualizations of reliability. Psychometrika 2005;70(1):123-33-33.