Validity of portfolio assessment: which qualities determine ratings?

Erik W Driessen,[1] Karlijn Overeem,[2] Jan van Tartwijk,[3] Cees P M van der Vleuten[1] & Arno M M Muijtjens[1]

CONTEXT The portfolio is becoming increasingly accepted as a valuable tool for learning and assessment. The validity of portfolio assessment, however, may suffer from bias due to irrelevant qualities, such as lay-out and writing style. We examined the possible effects of such qualities in a portfolio programme aimed at stimulating Year 1 medical students to reflect on their professional and personal development. In later curricular years, this portfolio is also used to judge clinical competence.

METHODS We developed an instrument, the Portfolio Analysis Scoring Inventory, to examine the impact of form and content aspects on portfolio assessment. The Inventory consists of 15 items derived from interviews with experienced mentors, the literature, and the criteria for reflective competence used in the regular portfolio assessment procedure. Forty portfolios, selected from 231 portfolios for which ratings from the regular assessment procedure were available, were rated by 2 researchers, independently, using the Inventory. Regression analysis was used to estimate the correlation between the ratings from the regular assessment and those resulting from the Inventory items.

RESULTS Inter-rater agreement ranged from 0.46 to 0.87. The strongest predictor of the variance in the regular ratings was 'quality of reflection' (R 0.80; R^2 66%). No further items accounted for a significant proportion of variance. Irrelevant items, such as writing style and lay-out, had negligible effects.

CONCLUSIONS The absence of an impact of irrelevant criteria appears to support the validity of the portfolio assessment procedure. Further studies should examine the portfolio's validity for the assessment of clinical competence.

KEYWORDS documentation/*standards/methods; education, medical, undergraduate/*methods; teaching/*methods; educational measurement /*standards/methods; observer variation; Netherlands.

Medical Education 2006; ***40****: 862–866*
doi:10.1111/j.1365-2929.2006.02550.x

INTRODUCTION

The strength of portfolios derives from their ability to offer rich and authentic evidence of learners' development and achievements. This makes them highly suitable not only for monitoring, but also for assessing learners' competence development. High validity is generally attributed to portfolio assessment.[1] Portfolios afford insight into learners' clinical competence through authentic evidential materials collected by learners in day-to-day practice over a prolonged period of time. Unfortunately, published studies on validity issues in relation to portfolios are rare. Indeed, the validity of portfolio assessment has been much less studied than the reliability of this type of assessment, particularly in medical education.[2] The few studies that we found in the literature lend some support to different aspects of portfolio validity: criterion and construct validity, predictive validity, and content validity.[3–5] Given that sound decisions require assessment that is both reliable and valid, the validity of portfolio assessment warrants further study.

Portfolio assessment is complex. This complexity is inherent to the open format of portfolios. An assessor

[1]Department of Educational Development and Research, Faculty of Medicine, University of Maastricht, Maastricht, The Netherlands
[2]St Radboud University Medical Centre, Nijmegen, The Netherlands
[3]ICLON Graduate School of Teaching, Leiden, the Netherlands

Correspondence: E W Driessen, Department of Educational Development and Research, University of Maastricht, PO Box 616, 6200 MD Maastricht, The Netherlands. Tel: 00 31 43 388 5774,
E-mail: e.driessen@educ.unimaas.nl

Overview

What is already known on this subject

Rich authentic evidence of learners' development makes the portfolio a valuable monitoring and assessment tool, but requires reliance on assessors' professional judgement, which may be affected by criteria with no relevance to the assessment purpose, such as writing style and lay-out.

What this study adds

Irrelevant criteria had negligible effects on the assessment of a portfolio aimed at stimulating reflection in students. Quality of reflection showed a correlation coefficient of 0.80 and was the strongest predictor of assessment results, explaining 66% of variance. The findings support the validity of this assessment procedure.

Suggestions for further research

Future research might examine the portfolio's validity in a larger sample and its validity in the assessment of clinical competencies.

has to judge portfolios that differ in content, size and, in many cases, structure. The richness and complexity of portfolios cannot be captured by analytic assessment criteria and detailed checklists can easily trivialise assessment.[6] That is why global (holistic) assessment methods characterised by strong reliance on assessors' professional judgements have been advocated for portfolio assessment.[7,8] A potential source of bias in such assessment procedures is that assessors may be tempted to let irrelevant qualities, such as the quality of the writing, the structure and the lay-out of the portfolio, sway their judgement. Presentation and students' personal characteristics have been shown to interfere with portfolio assessment, because they may mistakenly be interpreted as signs of competence in the area of interest.[1] For instance, in a study on portfolio assessment in teacher education by Quinlan, the analysis of thinking-aloud protocols revealed that competence ratings were influenced by what assessors already knew about students.[9] Heller *et al.*, who also used thinking-aloud protocols, stated: 'What is essential for maintaining valid scores is that raters be capable of consciously discriminating between relevant and irrelevant qualities during the rating judgement.'[10]

Because of the real risk that bias due to irrelevant portfolio qualities might compromise the validity of portfolio assessment, we designed a study to examine which criteria determined portfolio assessment of reflective competence. Our research questions were:

- Which criteria affect raters' judgements of students' reflective skills?
- Which criteria carry the most weight?

METHODS

Context

At Maastricht University, the Netherlands, portfolios are currently used to stimulate the development of students' academic and professional competencies over the course of the 6-year undergraduate medical curriculum.[11] Early in the curriculum, the portfolio's prime focus is on reflective skills, because reflection is regarded as a prerequisite for effective learning from experience. In this context, reflection is seen as a cyclic process. It starts with students analysing their learning experiences. In the next step, students distil learning objectives from the strengths and weaknesses emerging from the analysis. These learning objectives are then pursued by the students during subsequent experiences, at which point the cycle of analysing experience and generating and pursuing learning objectives starts afresh, to be repeated again and again. It is the role of the portfolio to invite students to record in words what they perceive as the strengths and weaknesses in their performance as well as their learning objectives, and whether and how they have attained those. In addition, students meet with their personal mentors at least twice a year to talk about the portfolio. The mentors try to steer students' efforts in fruitful directions. Instructions and guiding questions are another way to enhance students' reflective processes. The portfolio template offers some degree of structure and uniformity: students are expected to record the outcomes of reflecting on their performance in 4 professional roles: as a medical expert; as a health care professional; as a scholar, and as a person.[12] At the end of the year the portfolios are judged in a summative holistic assessment procedure based on the following (global) criteria:[8]

1 that the student's analyses of strengths and weaknesses of performance in the different roles are appropriate;
2 that the student has generated learning objectives that are clearly defined and feasible;
3 that the student has attained the learning objectives to a sufficient degree;
4 that the student provides appropriate evidence to support the analyses of strengths and weaknesses, and
5 that the portfolio contains all the required items and was handed in on time.

The mentors are trained in coaching students' portfolio and reflective skills at the start of mentorship. Just before the end-of-year assessment the mentors are trained in assessing the portfolios. The training consists of discussing and benchmarking portfolios from a previous year. The end-of-year assessment yields a rating of reflective competence as poor (fail), satisfactory or good. Mentors propose ratings and discuss these with the students, who are asked to express agreement or disagreement. Subsequently, each portfolio is assessed by another mentor and the final rating is determined by the assessment committee, which is composed of all the mentors. The assessment committee discusses only those portfolios that give rise to differences of opinion between mentor, student and/or other mentor.[8]

Instrument

In order to study the effects of 'irrelevant' qualities on portfolio ratings, we developed an instrument designated the Portfolio Quality Analysis Scoring Inventory. The 15 items on this Scoring Inventory are based on interviews conducted by the first author with the mentors of Year 1 medical students about criteria for portfolio content, structure and presentation, and the above-mentioned assessment criteria for reflective competence.[12] The items consist of propositions to be rated on a Likert scale from 1 = strongly disagree to 5 = strongly agree (Table 1). Scoring instructions explain when the different Likert scores are appropriate.

In a pilot of the Scoring Inventory, the second author and 1 of the mentors rated 10 portfolios using the Inventory. All 15 items were found to be relevant, but interrater agreement was unsatisfactory on 8 items. The wording of these items was revised to make them more accurate and precise. The scoring instructions were also refined.

Table 1 Interrater agreement on the Inventory items (significance $P < 0.005$; except for item 2 significance $P < 0.025$)

	Inventory items	Weighted kappas
1	This portfolio is among the best as regards lay-out	0.62
2	This portfolio is among the best as regards spelling and sentence structure	0.46
3	This portfolio is well structured: i.e. content is presented in the proper place, descriptions, analyses and learning objectives are easy to find	0.64
4	The portfolio is complete; i.e. no required components are missing	0.87
5	The student has looked critically at him/herself; i.e. indicates both strengths and weaknesses for the roles on which work was done	0.72
6	The analyses of strengths and weaknesses include a search for both internal and external explanations. The analysis is not limited to an enumeration of facts and/or situations	0.71
7	The analyses of strengths and weaknesses contain a sufficient number of different themes for each role	0.70
8	The student refers to evidence included in the portfolio in a systematic fashion; i.e. the evidence supports the analyses of strengths and weaknesses	0.63
9	The student has made a connection between extracurricular activities and and the development of his or her competencies	0.77
10	As for portfolio content, the student has done more than merely follow the guiding questions	0.62
11	The student refers to earlier versions of the portfolio (what went wrong, what went well this time and why, which statements did I make earlier)	0.62
12	The student has formulated logical (following from the analyses of strengths and weaknesses) and clearly defined learning objectives	0.70
13	The student explains how he or she wants to achieve the learning objectives	0.70
14	The student has attempted to show what he or she has undertaken to achieve the learning objectives	0.53
15	The student has expended more effort on the portfolio than was absolutely necessary	0.61

Table 2 Results of the regular portfolio assessment procedure for all Year 1 portfolios and the study sample

Rating	Year 1 ($n = 231$)	Sample from Year 1 ($n = 40$)
Poor	25 (10.82%)	5 (12.50%)
Satisfactory	124 (53.60%)	20 (50.00%)
Good	82 (35.50%)	15 (37.50%)

Procedure

We collected a stratified sample of 40 portfolios out of a total of 231 portfolios compiled by Year 1 students. These portfolios had already passed through the end-of-year assessment procedure described above. The sample was reasonably representative as regards the initial ratings and distribution across different mentors (Table 2). The students gave permission for the use of their portfolios for research purposes.

The second author and a teacher who was otherwise not involved in the study independently rated the 40 portfolios using the Portfolio Quality Analysis Rating Inventory. To enhance inter-rater reliability, the raters discussed the items after rating 2 portfolios. These portfolios were not included in the study. The raters were blinded to the results of the regular assessment procedure.

Data analysis

We used the mean Inventory score across assessors as an indicator of the quality of the portfolio. The reliability of this indicator was obtained by calculating the inter-rater agreement using weighted kappa statistics (using 'absolute error weights').[13,14]

We used stepwise multiple regression analysis to determine which of the Inventory items best predicted the ratings of reflective competence.

RESULTS

Inter-rater agreement was acceptable, ranging from 0.46 (moderate agreement) to 0.87 (excellent agreement) (Table 1).[15]

The standardised regression coefficients (beta) reflect the relative contributions of the independent variables (Table 3). Significance was only found for the item representing 'quality of reflection' (item 6). The bivariate correlation coefficient for this item and the regular rating was 0.80; thus quality of reflection explained 64% of the variance in the regular end-of-year ratings ($P < 0.000$). The standard error of estimate – the measure of residual variance – was 0.41.

Table 3 Regression coefficient (R) and explained variance (R^2) for the Inventory item on quality of reflection

Item	R	R^2	Standard errorof estimate	Significance
Quality of reflection	0.803	64%	0.405	0.000

After the entry of item 6, no further items accounted for a significant proportion of variance.

CONCLUSIONS

We examined which Inventory items affected portfolio assessment of reflective competence and which of these items carried the most weight. The results show that the ratings were primarily determined by quality of reflection. Quality of reflection was the only item to make a significant contribution (64%) to the explanation of the variance of the regular portfolio ratings. None of the other items of the Portfolio Quality Analysis Scoring Inventory made a substantial contribution.

The non-significant effects that we found for lay-out (item 1), spelling and grammar (item 2) and structure (item 3) as compared with quality of reflection (item 6) may be attributable to the mentoring and assessment training, which thus appears to positively affect the validity of the assessment procedure. Mentors and students meet twice a year and in those meetings mentors try to guide students' efforts in profitable directions. Mentors participate in assessment training just before the end-of-year assessment. During this training session, mentors are given approximately 20 minutes to independently assess fragments of sample portfolios from previous years. These assessments are then compared for benchmarking and the mentors explain which criteria were decisive in determining their judgements. In this way it becomes clear which criteria are relevant and mentors learn to discriminate between relevant and irrelevant criteria.[10]

Apart from the mentor training, the portfolio template may be another factor that explains our findings.

A limitation of this study is its relatively small sample. Further research on a larger sample may reveal more significant regression coefficients. Another limitation is that the portfolio focused exclusively on reflective competence and did not include general clinical competencies.

Although we did not explicitly investigate this, the study appears to provide some support for the reliability of the assessment items as well. We found rather high levels of agreement between 2 raters with regard to most items, not only the more objective ones, such as 'The portfolio is complete', but also more subjective items, such as 'The student has looked critically at him/herself'. Other studies on portfolio assessment suggested that inter-rater reliability is probably enhanced when criteria are discussed by (2) independent raters, as was done in this study.[16,17]

Because portfolio ratings were found to be associated with quality of reflection and not with aspects of presentation and writing style, we think we can conclude that the results support the validity of the global (holistic) assessment procedure for the assessment of reflective competence. Raters using the procedure appeared to be unaffected by irrelevant portfolio qualities in reaching their judgements.

Contributors: all authors contributed ideas to the paper and commented on all drafts. EWD designed and coordinates the portfolio programme in the Maastricht undergraduate medical curriculum.
Acknowledgements: we would like to thank Mereke Gorsira for revising the English language of this article. We would also like to thank the editor-in-chief of the Dutch medical education journal *Tijdschrift voor Medisch Onderwijs* for permission to use some of the data previously published in the journal: Overeem K, Driessen EW, Drenthe A, van Tartwijk J, van der Vleuten CPM. Kwaliteitscriteria voor portfoliobeoordeling van geneeskundestudenten. *Tijdschrift voor Medisch Onderwijs* 2005; 24(2):61–3.
Funding: none.
Conflicts of interest: none.
Ethical approval: not required.

REFERENCES

1 McMullan M, Endacott R, Gray MA, Jasper M, Miller CML, Scholes J, Webb C. Portfolios and assessment of competence: a review of the literature. *J Adv Nurs* 2003;**41**(3):283–94.
2 Pitts J, Coles C, Thomas P. Educational portfolios in the assessment of general practice trainers: reliability of assessors. *Med Educ* 1999;**33**:515–20.
3 Cadbury-Amyot C, Kim J, Palm R, Mills E, Noble E, Overman P. Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene programme. *J Dent Educ* 2003;**67**(9):991–1002.
4 Tillema H. Design and validity of a portfolio instrument for professional training. *Studies Educational Eval* 1998;**24**(3):265–74.
5 Valencia S, Au K. Portfolios across educational contexts: issues of evaluation, teacher development and system validity. *Educational Assessment* 1997;**4**(1):1–35.
6 Norman GR, van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ* 1991;**25**:119–26.
7 Ryan JM, Kuhs TM. Assessment of pre-service teachers and the use of portfolios. *Theory Pract* 1993;**32**:75–81.
8 Driessen EW, van der Vleuten CPM, Schuwirth L, van Tartwijk J, Vermunt JD. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ* 2005;**39**:214–20.
9 Quinlan KM. Inside the peer review process: how academics review a colleague's teaching portfolio. *Teach Teacher Educ* 2002;**18**:1035–49.
10 Heller J, Sheingold K, Myford C. Reasoning about evidence in portfolios: cognitive foundations for valid and reliable assessment. *Educational Assessment* 1998;**5**(1):5–40.
11 Driessen EW, van Tartwijk J, Vermunt JD, van der Vleuten CPM. Use of portfolios in early undergraduate medical training. *Med Teacher* 2003;**25**(1):18–23.
12 Driessen EW, van Tartwijk J, Overeem K, Vermunt JD, van der Vleuten CPM. Conditions for successful use of portfolios for reflection. *Med Educ* 2005;**39**:1230–5.
13 Cohen JA. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;**70**:213–20.
14 Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. *J Clin Epidemiol* 1993;**46**:1055–62.
15 Kianifard F. Evaluation of clinimetric scales: basic principles and methods. *Statistician* 1994;**43**:475–82.
16 Rees C, Sheard C. The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience. *Med Educ* 2004;**38**:138–44.
17 Pitts J, Coles C, Thomas P, Smith F. Enhancing reliability in portfolio assessment: discussions between assessors. *Med Teacher* 2002;**24**(2):197–01.

Received 19 September 2005; editorial comments to authors 13 December 2005; accepted for publication 14 March 2006