EDUCATION FOR HEALTH

ORIGINAL RESEARCH PAPER

Portfolio Assessment during Medical Internships: How to Obtain a Reliable and Feasible Assessment Procedure?

NRM Michels[1], EW Driessen[2], AMM Muijtjens[2], LF Van Gaal[1], LL Bossaert[1], BY De Winter[1]

[1]*Faculty of Medicine, University of Antwerp, Belgium*

[2]*Department of Educational Development and Research, Faculty of Medicine, University of Maastricht, Maastricht, The Netherlands*

Published: *December 2009*

Michels NRM, Driessen EW, Muijtjens AMM, Van Gaal LF, Bossaert LL, De Winter BY,

Portfolio Assessment during Medical Internships: How to Obtain a Reliable and Feasible Assessment Procedure?
***Education for Health*, Volume 22, Issue 3, 2009**

Available from: http://www.educationforhealth.net/

ABSTRACT

Background: A portfolio is used to mentor and assess students' clinical performance at the workplace. However, students and raters often perceive the portfolio as a time-consuming instrument.

Purpose: In this study, we investigated whether assessment during medical internship by a portfolio can combine reliability and feasibility.

Methods: The domain-oriented reliability of 61 double-rated portfolios was measured, using a generalisability analysis with portfolio tasks and raters as sources of variation in measuring the performance of a student.

Results: We obtained reliability (☐ coefficient) of 0.87 with this internship portfolio containing 15 double-rated tasks. The generalisability analysis showed that an acceptable level of reliability (☐ = 0.80) was maintained when the amount of portfolio tasks was decreased to 13 or 9 using one and two raters, respectively.

Conclusions: Our study shows that a portfolio can be a reliable method for the assessment of workplace learning. The possibility of reducing the amount of tasks or raters while maintaining a sufficient level of reliability suggests an increase in feasibility of portfolio use for both students and raters.

Keywords: Portfolio; medical education; assessment; clinical competences; workplace; internship; reliability; feasibility

Background

The assessment of clinical performance has received greater attention over the past few years. Clinical performance is defined as the combination and integration of different competences, such as knowledge, clinical skills, attitudes and professionalism. The assessment format should take into account the integral picture of the student (Miller, 1990). Subsequently, it is difficult to extrapolate traditional assessment theories (Norcini, 2005). A portfolio is seen as one of the potential instruments that can provide an acceptable judgment of clinical performance (Royal College of General Practitioners, 1993; Davis et al., 2001; Carraccio & Englander, 2004; Rees & Sheard, 2004; Driessen et al., 2006). The concept of a portfolio – as an assessment and coaching instrument – allows, on the one hand, evaluation of students on various competences and, on the other hand, assessment of the ongoing learning activity. The following definition of a portfolio, as "a collection of papers and other forms of evidence that learning has taken place" implies that it is possible to follow, remediate and assess the students through the different stages in their growth towards becoming a professional (David et al., 2001; Davis et al., 2001). In medical education, the portfolio is often used to assess professional development, mostly by including personal reflection tasks (Dornan et al., 2002; Duque et al., 2006). Overall, it remains crucial that the content of the portfolio is related to the purposes and roles of the portfolio within the educational program (Colbert et al., 2008).

Driessen et al. (2007) report acceptable interrater reliabilities for portfolio assessment for medical education which are in contrast with the results obtained in other domains such as the education of general practice trainers (Pitts et al., 1999) or primary educators (Koretz et al., 1994). In the same review, it is concluded that portfolios lose their utility and effectiveness if they ask for too much paperwork and time from students and faculty (Driessen et al., 2007). The problems of reliability and the time-consuming nature of the portfolio are also addressed by Colbert et al. (2008). As both reliability and feasibility are mandatory issues for the success of a portfolio, both aspects need to be considered (van der Vleuten & Schuwirth, 2005). To date, results regarding these two topics are often contradictory as it is generally accepted that reliability can be increased by using more raters, logically reducing the feasibility of portfolio assessment (Colbert et al., 2008; Driessen et al., 2007). Melville et al. (2004), for instance, report the need of four raters to achieve a reliability >0.8. Similarly, Burch & Seggie (2008) recognize the resource-intensive character of working with a portfolio. Their solution to increase the feasibility is to interview the students on a selection of the portfolio content. However, they did not investigate the effect of this procedure on the reliability of their portfolio, and the task burden for students remains. Another option to improve feasibility is creating a norm-based assessment, where students' performances will be compared with each other and will be ranked. This form of assessment, in contrast to criteria-based assessment in which an established set of criteria will be used to rate the content of the portfolio, will be less time-prohibitive (Davis et al., 2009).

Therefore, the aim of our study was twofold. First, we wanted to evaluate whether portfolio assessment during internships offers sufficient information for reliable judgment. Second, we investigated whether the feasibility for both students and raters could be improved. A generalisability analysis was performed, focussing on the domain-oriented reliability. Therefore the index of dependability was calculated, which quantifies the impact of the different sources of variation (students, tasks, raters) on the reliability and suggests adequate numbers of portfolio tasks.

Methods

Context:

In 1998, the curriculum of the University of Antwerp medical school (Belgium) was revised from a teacher-centred into a student-centred curriculum with an outcome-based approach. As it is known that assessment drives learning (Crossley et al., 2002a; van der Vleuten & Schuwirth, 2005), the implementation of an assessment program in the revised curriculum attracted considerable attention. In the sixth year of medical school, a revised internship evaluation document (evaluation through observation) was introduced next to portfolio assessment (evaluation through written tasks), both assessing different workplace competences. The combination of different assessment tools helps to evaluate students in a reliable and valid way (van der Vleuten & Schuwirth, 2005) as they can measure different competences, judge students in a different way and demand more or less time from students and raters. In this regard, our faculty agreed that the advantage of a portfolio to measure different competences and their integration, next to the evaluation at the workplace, increases the educational value and is therefore worth the well-known time-investment.

The development of the portfolio was based on a blueprint with three dimensions: 1) competence type, 11 competences deduced from the three-circle model of Harden and colleagues (1999); 2) competence level, according to the pyramid of Miller (1990); and 3) assessment tools.

During their fulltime (12 months) internship in year 6, medical students rotated through nine different disciplines in the hospital: internal medicine; surgery; paediatrics; obstetrics and gynaecology; (adult and/or adolescent) psychiatry; ophthalmology; otorhinolaryngology; dermatology; and general practice. The portfolio consisted of 15 tasks that were clustered into four categories (see Table 1): case reports on patient encounters; scientific medical presentations; self–reflections; and tasks specifically linked to a discipline. More specifically, students were asked to write seven case reports linked to the following disciplines: internal medicine; surgery; paediatrics; (adult and/or adolescent) psychiatry; ophthalmology; otorhinolaryngology; and dermatology. Additionally, they had to complete two presentations (internal medicine and surgery), three self-reflections (topics linked to internal medicine, paediatrics and psychiatry) and three specific tasks, consisting of a surgery report, an obstetric report and a general practice task.

Students were encouraged to accentuate their personality in the tasks according to their own experiences, interests, knowledge and scientific background and to justify their portfolio content and choices. Students and raters were provided with a description of each type of task in the portfolio and guidelines concerning the purpose of the portfolio as well as the assessment procedure.

Portfolio assessment procedure:

We chose a double rating system based on an 8-point global rating scale using trained evaluators. On the one hand, portfolios were randomly divided among the lecturers of the Skills Lab (n=7); they evaluated and rated each task and obtained a global view of the portfolio of that particular student. On the other hand, clinicians of different disciplines (n=10) assessed the tasks linked to their discipline. As we took the option of a criteria-based assessment, specific criteria per type of task were listed and communicated to students and raters (see Table 1). For example, when writing self-reflections, students had to complete all the phases of Korthagen's cycle (1985). This cycle or ALACT-model structures reflection tasks on practice-related events, including five subsequent phases: the **A**ction or experience, **L**ooking Back on own feelings and thoughts, the **A**wareness of essential aspects, **C**reating alternative methods and **T**rialling in a new situation. In addition, more general criteria were applied to all tasks; these criteria included medical and scientific correctness, as well as a systematic and logical approach to a task. The diversity of the topics chosen and the personal way of composing the portfolio were also taken into account. Layout and writing style were not assessed but a certain level of order, clear structure and appropriate language use were required.

All raters were trained and had the expertise to assess the portfolio tasks according to the specified criteria. Per task, an 8-point global rating scale was used, ranging from 1 = bad fail to 8 = excellent (Davis et al., 2001). The final score for each task in the portfolio was the mean of the two

raters' scores. The global portfolio score was obtained by averaging all 15 task scores, assuming equal weight for each task in the portfolio. Pass/fail decisions were based on comparison of the global score with a standard absolute norm.

Table 1: Portfolio tasks clustered into four categories, with descriptions of their criteria

A. Case report *(internal medicine, surgery, paediatrics, psychiatry, ophthalmology, otorhinolaryngology, dermatology)*
• choice of the topic/case
• medical and scientific correctness
• systematic approach and completeness
  ○ history-taking
  ○ clinical investigation
  ○ differential diagnosis
  ○ technical investigation
  ○ diagnosis
  ○ therapy
  ○ conclusion/literature
• motivation/learning needs and plans
	B. Presentation *(internal medicine, surgery)*
	• choice of the topic/case
	• medical and scientific correctness
	• clear presentation (PPT, slides)
	• literature search/own conclusions

C. Self-reflection *(internal medicine, paediatrics, psychiatry)*
• choice of the topic, relevancy
• all phases of cycle of Korthagen
  ○ **A**ction: description of **own** experience/critical event
  ○ **L**ooking Back: analysis of **own** feelings/thoughts
  ○ **A**wareness of essential aspects
  ○ **C**reating alternative methods
  ○ **T**rial
• personal point of view
	D. Tasks specifically linked to a discipline *(surgery report, obstetric report, general practice task)*
	• choice of the topic
	• medical and scientific correctness
	• systematic approach and completeness
	• conformity to guidelines of task

Data analysis:

The data consist of the final scores of 15 portfolio tasks for all 6th year medical students (n=61) in the academic year 2005–2006. The domain-oriented reliability (reliability of the student's score with respect to the competence domain) was calculated using generalisability theory. There are several methods for measuring reliability. When assessing complex behaviours and more subjective tasks, a generalisability analysis is advised (Crossley et al., 2002b). First, it is an elegant way to measure reliability and the variance of the components as sources of bias (Downing, 2004). Second, it provides scientific advice as to the minimum amount of tasks required to obtain a sufficient level of reliability.

A generalisability analysis was performed (model: Rater: Student x Task). As we were interested in the reliability of the student's score with respect to the competence domain (absolute interpretation of scores), the dependability index (□ coefficient) was also calculated (Brennan, 2001). The variance components for Student, Task, Student-Task interaction and Rater within Student-Task interaction were estimated in a G-study (Crossley et al., 2007). Finally, based on these estimates, in a D-study (decision study) the reliability for a hypothetical number of Tasks and Raters was calculated (Crick & Brennan, 1983).

To investigate the time-consuming nature of the portfolio, all the Skills Lab raters were asked to record the time needed to evaluate the portfolios. In addition – 12 students - selected randomly, were asked retrospectively the average time needed to complete a single portfolio task. These results are presented as mean ± standard deviation.

Results

Details on the scores of the different tasks are provided in Table 2, with mean scores ranging between 5.70 ± 0.89 and 6.59 ± 0.79 (scale 1-8; mean ± standard deviation). The total portfolio score was 6.26 ± 0.59, based on a maximum score of 8. The G-study resulted in variance components for Student, Task, Student-Task interaction and Rater within Student-Task interaction of, respectively, 24%, 5%, 24% and 47% of the total variance. Hence, the difficulty variation over Tasks was small (5%), but there was a considerable within-student variation over Tasks (24%), which is specific for each student. This implies that a student does not perform each of the tasks at the same level. This is generally indicated as 'case-specificity', which is often found to be a major source of measurement variance in assessment (van der Vleuten & Swanson, 1990). The largest variance component was due to the Rater within Student-Task interaction (47%), indicating that there is a considerable difference between the scores of the two different groups of raters. The index of dependability for 15 tasks was 0.87 (Table 3). The D-study (hypothetical generalisability analysis) showed that for the currently used double-rating procedure, a reliability of 0.8 would be obtained with 9 tasks, being 60% of the original portfolio content (Table 3). If using a single-rating procedure, 13 portfolio tasks would be required.

Table 2: Task scores and total portfolio score on an 8-point global rating scale (n = 61 students)

	mean of the scores	standard deviation
internal medicine case report	6.16	0.76
internal medicine presentation	6.25	0.83
surgery case report	6.30	1.20
surgery presentation	6.59	0.79
surgery report	6.47	0.79
gynaecology case and obstetric report	6.51	0.68
paediatrics case report	6.35	0.84
psychiatry case report	6.56	0.75
ophtalmology case report	6.50	0.81
otorhinolaryngology case report	6.45	0.87
dermatology case report	5.70	0.89
general practice	6.04	1.20
self-reflection internal medicine	5.89	1.14
self-reflection paediatrics	6.07	1.11
self-reflection psychiatry	6.15	1.10
total portfolio	6.26	0.59

Table 3: Dependability index (domain-oriented generalisability) for varying numbers of raters (nR) and tasks (nT), showing a dependability index of 0.87 (☐ coefficient) for 15 double-rated tasks (arrow 1). A ☐ coefficient of 0.80 can be obtained for 13 single-rated tasks (arrow 2) or 9 double-rated tasks (arrow 3)

nR \ nT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.24	0.39	0.49	0.56	0.61	0.65	0.69	0.72	0.74	0.76	0.78	0.79	**0.80**	0.82	0.83	0.84
2	0.31	0.48	0.58	0.65	0.70	0.73	0.76	0.79	**0.80**	0.82	0.83	0.85	0.86	0.87	**0.87**	0.88

3 2 1

The mean time raters needed to evaluate a portfolio containing 15 tasks was 181 ± 48 minutes. Students needed approximately 7.6 ± 3.5 hours to execute a single portfolio task, resulting in approximately 114 hours of student workload for the global portfolio.

Discussion

In this study, we investigated the reliability of portfolio assessment during internships in order to find out whether the feasibility of the procedure could be improved by reducing the task burden while retaining a level of reliability which is sufficient for summative assessment. The results of our study indicate that a portfolio assessment procedure, including multiple and varied tasks and a double-rating system, can be used for summative assessment. The overall dependability index of 0.87 confirms acceptable reliability (Magnusson, 1967). Further, the D-study (the hypothetical part of the generalisability analysis) shows that a reliability of 0.8 or higher can be maintained with nine tasks, amounting to a reduction of the portfolio task burden of 40%. Using a single- instead of double-rating system would imply a 50% reduction regarding time and effort for raters. However, the allowed reduction of tasks would be much smaller: 13 tasks would be required to maintain a reliability of 0.8, hence, a reduction of only 13%.

Developing a portfolio assessment, like any other assessment format, is a balance between creating an effective, reliable learning and assessment tool and considering the practical context-related factors (Crossley et al., 2002a; van der Vleuten & Schuwirth, 2005). We showed that a generalisability analysis can be helpful, next to educational and managerial considerations, in improving the feasibility for both students and raters.

To date, few studies have addressed methods to improve feasibility (Driessen et al., 2007). To obtain reliable assessment scores, the use of an established set of criteria can be advised. Implementing criteria-based assessment in favour of norm-based assessment will have a positive impact on the interrater reliability, even though feasibility can be decreased. Furthermore, reliability can be increased by sufficiently sampling across the content and the competences to be evaluated and/or to involve a sufficient number of raters in the assessment process (van der Vleuten & Schuwirth, 2005). However, by increasing the number of tasks or raters, feasibility and acceptability can be influenced negatively. Portfolios are at risk of becoming a bulk of paperwork and cease to be effective (Davis et al., 2009).

Assessment by one instead of two raters per task would reduce the time-investment for the group of raters as for one group the time investment will be reduced to zero. However, we think this is not the ideal solution in our setting for two reasons. First, the results of the generalisability analysis show that the corresponding reduction of task burden for students is relatively small. To maintain a reliability of 0.8 in a one-rater scenario, the mean time investment per portfolio per student can be reduced from 114 hours (15 tasks) to 99 hours (13 tasks), representing a reduction of 15 hours per student. Yet, a reliability level of 0.8 can be maintained in a double-rater scenario by 9 instead of 15 tasks, corresponding to 68 hours of student time - hence, a reduction of 46 hours per student. Second, from the perspective of the validity of the assessment, we value the judgment of both groups of raters (clinicians and lecturers of the Skills Lab) as independent rater groups with their own perspectives and expertise (clinical versus educational emphasis). In addition, the considerable variance in the Rater within Student-Task interaction reflects a variation in assessing students and their tasks within the total group of raters, and thus also supports maintaining the double-rating, each from their point of view starting from the same evaluation criteria.

Additionally, we should be aware that decreasing the number of tasks in the portfolio could have consequences for the validity of the assessment as learning outcomes could be compromised. To maintain a valid instrument, an adequate sampling across all required competences is essential. Therefore, we have to keep in mind the initial blueprint with respect to the intended focus and purposes of the portfolio. Our portfolio consisted of seven case reports, two presentations, three self-reflections and three special tasks. Each task received equal weight, thereby favouring the influence of the task type case reports on the total score. Therefore, it seems advisable in our portfolio format to reduce the number of the

different tasks according to the results obtained in this generalisability analysis resulting in a similar weight of the four quadrants (Table 1). Practically, this would point to a greater reduction in case reports than in the amount of self-reflections, presentations and specific tasks in an attempt to maintain a valid instrument.

This brings us to an important limitation of our study: reliability is a necessary but not sufficient condition for validity (Crossley et al., 2002a; Downing, 2004). By the use of a competence-based blueprint and the implementation of most of the suggested characteristics and rules such as clear assessment criteria, guidelines and experienced raters in the development of the portfolio as assessment procedure, we addressed a certain level of validity (Downing, 2003; Driessen et al., 2006; Driessen et al., 2007). However, to strengthen the validity of our portfolio assessment procedure, research needs to be performed on the renewed portfolio to ascertain that we are measuring the competences we aim to measure with our tasks (content validity). A Delphi study will be performed to link the targeted assessed competences to the content of the portfolio. Another limitation is the fact that we only investigated portfolio use in the context of students' internships in a single medical school.

Conclusions

Motivated by critical reflection and by the need for a scientific approach to portfolios as an assessment tool, a generalisability analysis was performed at the University of Antwerp medical school. Results showed that an assessment procedure with a double-rating system obtained a high reliability with 15 tasks. Additionally, the generalisability analysis demonstrated that it is possible to maintain an acceptable level of reliability while reducing the amount of tasks by 40% in the two-rater system, thus enhancing the feasibility of the portfolio assessment.

Aknowledgements

The authors would like to thank the raters, both the clinicians and the lecturers of the Skills Lab, for their participation and effort.

References

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Burch, V. C., & Seggie, J. L. (2008). Use of a structured interview to assess portfolio-based learning. *Medical Education, 42*, 894-900.

Carraccio, C., & Englander, R. (2004). Evaluating competence using a portfolio: A literature review and Web-based application to the ACGME competencies. *Teaching and Learning in Medicine, 16(4)*, 381-387.

Colbert, C. Y., Ownby, A. R., & Butler, P. M. (2008). A review of portfolio use in residency programs and considerations before implementation. *Teaching and Learning in Medicine, 20*(4), 340-345.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system (ACT Technical Bulletin No. 43)* . Iowa City, IA: ACT, Inc.

Crossley, J., Humphris, G., & Jolly, B. (2002a). Assessing health professionals. *Medical Education, 36*(9), 800-804.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002b). Generalisability: A key to unlock professional assessment. *Medical Education, 36*(10), 972-978.

Crossley, J., Russell, J., Jolly, B., Ricketts, C., Roberts, C., Schuwirth, L., et al. (2007). "I'm pickin' up good regressions': The governance of generalisability analyses. *Medical Education, 41*(10), 926-934.

David, M. F. B., Davis, M. H., Harden, R. M., Howie, P. W., Ker, J., & Pippard, M. J. (2001). AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Medical Teacher, 23*(6), 535-551.

Davis, M. H., Ben-David, M. F., Harden, R. M., Howie, P., Ker, J., McGhee, C., et al. (2001). Portfolio assessment in medical students' final examinations. *Medical Teacher, 23*(4), 357-366.

Davis, M. H., Ponnamperuma, G. G., & Ker, J. J. (2009). Student perceptions of a portfolio assessment process. *Medical Education, 43*(1), 89-98.

Dornan, T., Carroll, C., & Parboosingh, J. (2002). An electronic learning portfolio for reflective continuing professional development. *Medical Education, 36*(8), 767-769.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. Medical Education, 38(9), 1006-1012.

Driessen, E. W., Overeem, K., van Tartwijk, J., van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2006). Validity of portfolio assessment: Which qualities determine ratings? *Medical Education, 40*(9), 862-866.

Driessen, E., van Tartwijk, J., van der Vleuten, C., & Wass, V. (2007). Portfolios in medical education: Why do they meet with mixed success? A systematic review. *Medical Education, 41*(12), 1224-1233.

Duque, G., Finkelstein, A., Roberts, A., Tabatabai, D., Gold, S. L., & Winer, L. R. (2006). Learning while evaluating: The use of an electronic evaluation portfolio in a geriatric medicine clerkship. *BMC Medical Education, 6*(1), 4.

Harden, R. M., Crosby, J. R., Davis, M. H., & Friedman, M. (1999). AMEE Guide No. 14: Outcome-based education: Part 5 - From competency to meta-competency: A model for the specification of learning outcomes. *Medical Teacher, 21*(6), 546-552.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5-16.

Korthagen, F. A. J. (1985). Reflective teaching and preservice teacher education in the Netherlands. *Journal of Teacher Education, 36*(5), 11-15.

Magnusson, D. (1967). *Test Theory*. Stockholm: Addison-Wesley.

Melville, C., Rees, M., Brookfield, D., & Anderson, J. (2004). Portfolios for assessment of paediatric specialist registrars. *Medical Education, 38*(10), 1117-1125.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*, S63-S67.

Norcini, J. J. (2005). Current perspectives in assessment: The assessment of performance at work. *Medical Education, 39*, 880-889.

Pitts, J., Coles, C., & Thomas, P. (1999). Educational portfolios in the assessment of general practice trainers: Reliability of assessors. *Medical Education, 33*(7), 515-520.

Rees, C., & Sheard, C. (2004). Undergraduate medical students' views about a reflective portfolio assessment of their communication skills learning. *Medical Education, 38*(2), 125-128.

Royal College of General Practitioners. (1993). *Portfolio-based Learning in General Practice: Report of a Working Group on Higher Professional Education: December 1993*. Unpublished manuscript.

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58-76.

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317.