# The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study

E Driessen,[1] C van der Vleuten,[1] L Schuwirth,[1] J van Tartwijk[2] & J Vermunt[3]

AIM Because it deals with qualitative information, portfolio assessment inevitably involves some degree of subjectivity. The use of stricter assessment criteria or more structured and prescribed content would improve interrater reliability, but would obliterate the essence of portfolio assessment in terms of flexibility, personal orientation and authenticity. We resolved this dilemma by using qualitative research criteria as opposed to reliability in the evaluation of portfolio assessment.

METHODOLOGY/RESEARCH DESIGN Five qualitative research strategies were used to achieve credibility and dependability of assessment: triangulation, prolonged engagement, member checking, audit trail and dependability audit. Mentors read portfolios at least twice during the year, providing feedback and guidance (prolonged engagement). Their recommendation for the end-of-year grade was discussed with the student (member checking) and submitted to a member of the portfolio committee. Information from different sources was combined (triangulation). Portfolios causing persistent disagreement were submitted to the full portfolio assessment committee. Quality assurance procedures with external auditors were used (dependability audit) and the assessment process was thoroughly documented (audit trail).

RESULTS A total of 233 portfolios were assessed. Students and mentors disagreed on 7 (3%) portfolios and 9 portfolios were submitted to the full committee. The final decision on 29 (12%) portfolios differed from the mentor's recommendation.

CONCLUSION We think we have devised an assessment procedure that safeguards the characteristics of portfolio assessment, with credibility and dependability of assessment built into the judgement procedure. Further support for credibility and dependability might be sought by means of a study involving different assessment committees.

KEYWORDS education, medical, undergraduate/*methods; educational measurement/*methods; curriculum/ standards; reproducibility of results; clinical competence/*standards; students, medical/*psychology; mentors.

*Medical Education 2005; **39**: 214–220*
doi:10.1111/j.1365-2929.2004.02059.x

## INTRODUCTION

The use of portfolios as an assessment tool has gained rapid popularity. As has happened with many assessment instruments, the term 'portfolio' has become a container concept covering a diversity of methods.[1,2] At the heart of every portfolio is information collected in evidence of the owner's learning process and/or competence levels. The evidence is often organised by competencies and may be supplemented with reflections on educational achievement and personal and professional development.[3] Portfolios were primarily introduced to assess performance in authentic contexts and encourage learners to reflect on their performance.[4] When portfolios are used for summative rather than formative assessment, the psychometric

[1]Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands
[2]ICLON, Leiden University, Leiden, The Netherlands
[3]IVLOS Institute of Education, Utrecht University, Utrecht, The Netherlands

*Correspondence*: Erik Driessen, Department of Educational Development and Research, PO Box 616, 6200 MD Maastricht, The Netherlands. Tel: 00 31 43 388 5774; Fax: 00 31 43 388 5779; E-mail: e.driessen@educ.unimaas.nl


 

## Overview

**What is already known on this subject**

Qualitative and holistic approaches to assessment need not be disqualified because of their inherent subjectivity and consequent lack of reliability.

**What this study adds**

Credibility and dependability of an assessment procedure can be built into the judgement process. These in-built procedures include intermittent feedback cycles, involvement of relevant resource persons, including the student, sequential information gathering until saturation of information is reached, and quality assessment procedures.

**Suggestions for further research**

Further support for this assessment procedure might be sought by means of a study involving different assessment committees.

qualities must meet stringent requirements, particularly in terms of reliability. What evidence the scant studies on the reliability of portfolios have revealed is cause for concern, as most studies have reported moderate to low (interrater) reliability.[3,5,6] The inevitable conclusion is that extreme caution is warranted when portfolios are used for summative purposes.[7] A case in point is the study carried out in the state of Vermont on the reliability of their large scale portfolio assessment programme in primary education.[8] In response to the low reliability scores (interrater [Spearman] correlations between 0·45 and 0·65), the Vermont Department of Education restricted the reporting of portfolio scores. The Vermont case attracted substantial attention and led to increased standardisation of portfolio assessment.

There are several strategies for improving interrater reliability in portfolio assessment:

1 standardisation: for example, by structuring content and restricting the number of admissible sources of evidence;
2 rater training and the structuring of judgement through checklists with strict performance criteria, and
3 using large numbers of raters to average out any rater effects.

The disadvantage of the first 2 strategies is that they jeopardise validity. Portfolios are valuable largely because of the richness of the information they supply. They enable students to present documentation of their personal, authentic, educational experiences and experiences in real practice. Standardising those experiences would inevitably detract from the portfolio's educational value. Training of raters and shared rater experiences would improve interrater reliability, but studies on the objective structured clinical examination have warned against unrealistically high expectations of rater training, even with well defined instruments.[9] The lesson that detailed checklists can easily trivialise assessment has also been learned in other assessment domains.[10] Increasing the number of raters would be an effective strategy, were it not for practical constraints, such as the time-consuming nature of portfolio judgement. In summary, portfolio assessment appears to be caught between the 2 classic evils of poor reliability and poor validity. This begs the question of how to achieve sufficient reliability of qualitative and subjective judgements for summative purposes without falling into the trap of 'corruption of portfolios for testing purposes'.[11]

Common misconceptions pervading this discussion are that subjectivity equals unreliability and that objectivity equals reliability. This is not universally true: objective examinations may be unreliable (cf. a single-item, multiple-choice examination) and – more importantly – subjective judgements can be reliable provided an adequate number of different judgements are collected and collated.[12] In any formal assessment procedure a fair decision must optimally reflect the demonstrated competence. This implies that assessments must be comparable across candidates, with minimisation of bias and error. Psychometrically, this means large sample sizes and structured (i.e. objective) assessments. In this paper we will present a qualitative approach to portfolio assessment that can enhance reliability without taking recourse to large samples and rigid structuring. Before looking at reliability from the perspective of qualitative research criteria, we will address some analogies between concepts. Some of the concepts underlying internal validity and reliability have pendants in credibility (cf. internal validity) and

dependability (cf. reliability) as used in qualitative research.[13] To assess the trustworthiness of qualitative data, Lincoln and Guba have systematically replaced traditional criteria by a set of parallel methodological criteria. A central criterion is credibility, which relates to the truth value within the findings so that they are both believable and supported by the evidence provided.[13] A number of methodological strategies have been suggested to ensure credibility and dependability.[14] The following 3 strategies are important for reaching credibility: *triangulation* (combining different information sources); *prolonged engagement* (sufficient time investment by the researcher), and *member checking* (testing the data with the members of the group from which they were collected). The strategies for realising dependability – the pendant of reliability – involve establishing an *audit trail* (i.e. documentation of the assessment process to enable external checks) and carrying out a *dependability audit* (i.e. quality assessment procedures with an external auditor).

These strategies can also be used to achieve credibility and dependability in educational assessment.[15] For example, Norcini and Shea use the concept of credibility as a criterion against which methods of standard setting can be evaluated.[16] In this approach, standard setters should have the proper qualifications (*prolonged engagement*), many standard setters should be involved and the judgements of other credible groups should be included (*triangulation*).[16] Credibility strategies are especially useful for constructing an assessment procedure, while dependability strategies can be used to monitor the assessment procedure. We conducted a case study among 237 Year 1 medical students to explore how the concepts of credibility and dependability can be applied to portfolio assessment. We addressed the question as to how such qualitative research strategies can be used in a portfolio assessment procedure to ensure reliable and valid judgement.

## CONTEXT OF THE STUDY

This case study explored the portfolio assessment procedure used in Year 1 of the undergraduate medical curriculum at Maastricht University, the Netherlands. The structure of the portfolio was provided by 4 different roles of a doctor: medical expert, scientist, health care worker and person. Global criteria were devised for each role and students had to collect evidence demonstrating that by the end of the year they had met those criteria. The students were mentored by medical school staff.

At the beginning of the academic year the portfolio system was introduced and students carried out some portfolio exercises. In the portfolio students had to present an analysis of their personal strengths and weaknesses in relation to the 4 roles of a doctor. These reflections had to be backed up by evidence, such as feedback from evaluations and tests and completed assignments. Students were also required to draw up a learning plan for the next period. Halfway through the academic year the students submitted their portfolios to their mentors, who gave feedback. In a progress meeting student and mentor discussed the portfolio and the student's competence development regarding the 4 roles. It was assumed that the student would adjust the portfolio in accordance with the feedback received. At the end of the academic year this cycle of submission, feedback and adjustment was repeated. This portfolio format has been described in greater detail elsewhere.[17]

### Formative and summative assessment

The purpose of the Year 1 portfolio was primarily *formative*. It was intended to promote feedback as part of the assessment programme and help students monitor their competence development and develop reflective, planning and remediation skills.

The portfolio also served a *summative* purpose. This was considered desirable for 2 reasons:

1 because experience has taught that purely *formative* assessment tends to lose momentum and after some time a new impetus is needed to steer student learning into the desired direction,[18] and
2 because portfolio assessment offers a unique opportunity to identify students who are lagging behind in professional progress and who show insufficient ability to reflect, plan and/or take remedial action. As *summative* assessment implies the possibility of failing and students who fail the portfolio may ultimately face expulsion from medical school, it will be clear that the portfolio is a high stakes assessment and fair decisions and maximum prevention of decision errors are of the essence.

In summary, the portfolio assessment procedure in the case study was designed to strike a delicate balance between *formative* and *summative* evaluation, seeking the best possible mix of benefits from both approaches.[19] We will describe how we combined *formative* and *summative* assessment in a single procedure.

## THE PORTFOLIO ASSESSMENT PROCEDURE

All portfolios ($n = 237$) were judged at the end of the academic year and given a grade of fail, pass or distinction. The following, rather global, criteria were used to assess the quality of the portfolios:

- the quality of the analyses of strengths and weaknesses;
- the quality of the evidence;
- the extent to which the evidence reflected the analyses of strengths and weaknesses;
- the clarity and feasibility of the learning objectives, and
- the extent to which the learning objectives were achieved.

These criteria express the following steps in the portfolio cycle: reflect on competence development; sample evidence; link evidence to reflection; formulate learning objectives, and develop competence. Obviously, such broad criteria necessitate extensive input from the judges in the assessment process.

Assessment occurred in all phases of the portfolio process:

1 during the compilation of the portfolio in regular meetings of mentor and student;
2 in the end-of-year meeting when mentor and student recommended the final grade, and
3 after submission of the portfolio to the portfolio assessment committee (PAC) for final grading.

In all 3 phases procedures and precautions had to be in place to ensure a credible assessment process.

### Compiling the portfolio

Over the course of the year the students discussed their progress as documented in the portfolio in at least 2 sessions with their mentor, who provided constructive oral and written narrative feedback.

The mentor's combined role of supervisor and assessor can be a difficult, albeit not impossible, task. A classic example of a situation involving a similar role combination is the relationship between supervisor and PhD student. The supervisor has to coach and encourage the student to put his or her best efforts into the dissertation, while at the same time both supervisor and student have to prevent submission of an unsatisfactory dissertation to an external assessment panel. In our case study, this analogy has been particularly useful in the training sessions for the mentors held at the beginning of the academic year. These sessions were supplemented by intervision sessions, in which the mentors shared experiences and information. The purpose of this approach was to support the mentors in their difficult double role and build a sound foundation for feedback to the students. Another advantage of regular feedback is that it prevents students being disappointed by an unexpected, negative recommendation made by a mentor to the portfolio committee.

### Recommendation by mentor and student

In their final meeting of the academic year the student and the mentor discussed the mentor's well motivated recommendation to the assessment committee concerning the grading of the portfolio. When student and mentor agreed on the grade, the student signed the recommendation. The student did not sign if there was disagreement, which the student indicated on the assessment form. Subsequently, the portfolio was submitted to the committee.

### Portfolio assessment committee

The final step of the assessment procedure comprised a sequential judgement procedure by the assessment committee. As it is the mentors who have first-hand experience with the portfolios, it was decided that the assessment committee should be composed of the 13 Year 1 mentors. The committee members did not grade the portfolios of the students they had mentored. Because judging a portfolio is time-consuming, the assessment procedure was designed for maximum efficiency. Judgements of the full committee were only required if the available information was not unanimous. Figure 1 presents the assessment procedure in a flowchart.

The flowchart shows the number of portfolios remaining to be judged after each step in the decision process. A total of 233 portfolios were submitted to the assessment committee at the end of the academic year 2001–02. Firstly, the portfolios on which student and mentor agreed were rated by a single committee member, who did not study the portfolio in any great detail, but typically scanned the work of the student and mentor and checked whether all procedures had been followed correctly. Only if the rater had any doubts was the portfolio examined further. When rater, mentor and student agreed on the grading, the

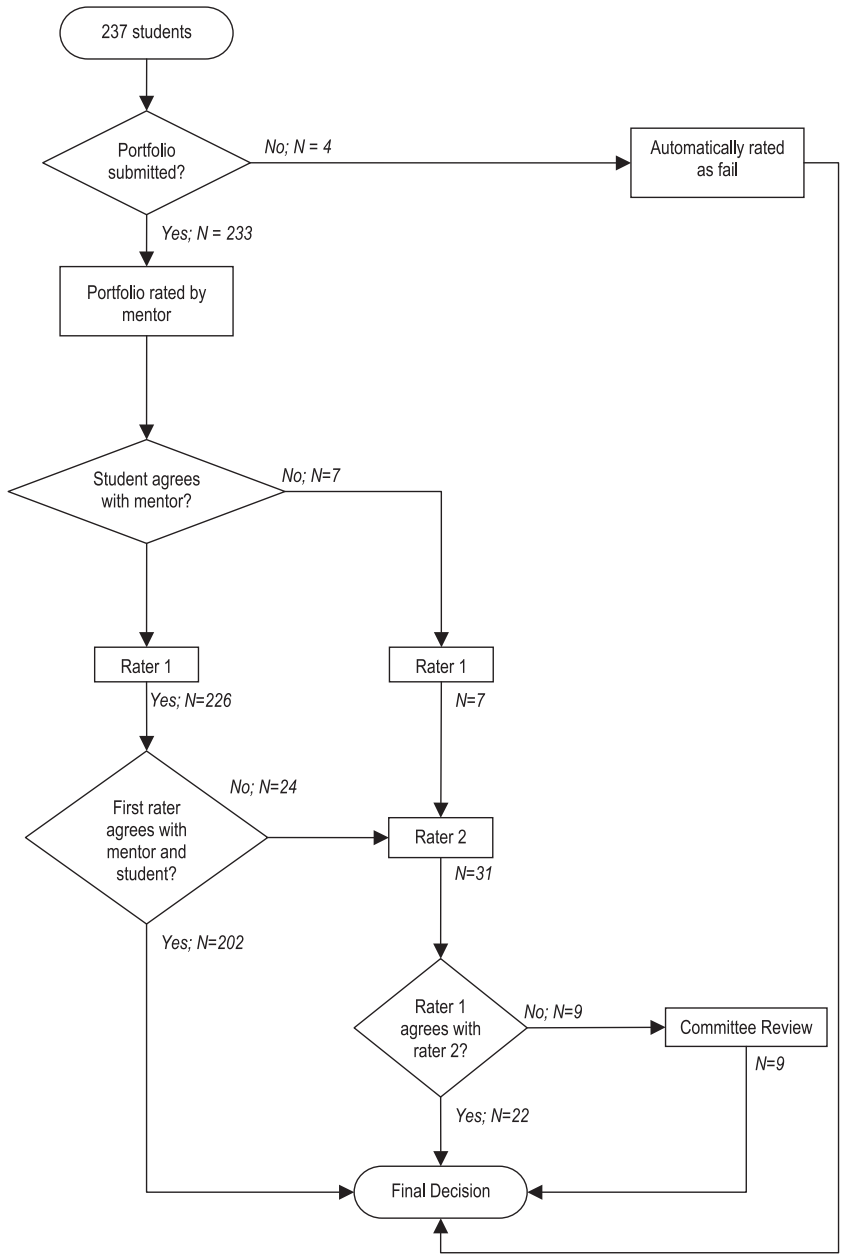


**Figure 1** Flow chart of the judgement procedure of the portfolio assessment committee.

recommendation became the final decision. The fact that this parsimonious route proved feasible for the majority of portfolios ($n = 202$; 85%) may be taken as an indication that the assessment procedure, with its mentor–student meetings, was satisfactory. If the rater did not agree with student and mentor, a second independent rater judged the portfolio. If the 2 raters agreed, their judgement became the final decision. If they disagreed, the portfolio was submitted to the full committee.

Students and mentors disagreed about the grading of only 7 (3%) portfolios. These portfolios were judged by 2 raters independently of one another. When the 2 raters agreed, their rating was regarded as the final decision. If they disagreed, the portfolio was submitted to the full committee.

Only 9 portfolios (4%) were discussed in the full committee meeting. These portfolios were reviewed individually, with mentors and raters presenting their arguments. The final decision was based on consensus among the committee members, excluding the student's mentor. As the mentor is first and foremost a supervisor, rather than an assessor, the mentor had no vote in the final decision. In all, 29 (12%) final

decisions differed from the original recommendation. Nine students failed, 147 received a pass and 81 were given a distinction.

A total of 226 portfolios (96%) were graded without being reviewed by the full committee. The entire procedure was completed in the relatively short time of 42 hours (i.e. 11 minutes per portfolio), with the committee meeting lasting 1 hour. The participants did not perceive the process as particularly stressful.

## DISCUSSION

This case study demonstrates the feasibility of a qualitative approach to achieve reliable summative judgement using an inherently complex and non-standardised assessment instrument, which relies on holistic professional judgement. We incorporated some procedural safeguards into the assessment process to achieve maximum credibility of the decisions.[13–15] Essential elements in the assessment process were:

- feedback cycles, incorporated into the mentoring process during the compilation of the portfolio to ensure that the mentor's final recommendation did not come as a(n) – unpleasant – surprise to the student; this element relates to the credibility strategies of prolonged engagement and member checking;
- maintaining a careful balance between the mentor's roles of coach and assessor, ensuring that the person who knew the student best provided the most relevant information while at the same time minimising any damaging effect to the mentor–student relationship; this relates to the credibility strategy of prolonged engagement;
- student involvement in the decision process to ensure commitment on the part of the student and allow the student to communicate a different point of view to that of the mentor; this relates to the credibility strategy of member checking, and
- a sequential judgement procedure in which conflicting information necessitated more information gathering, ensuring efficient use of resources by reserving efforts to achieve more reliable judgement in cases where this was absolutely necessary. As a result, more resources (i.e. mentor time) were available for coaching students, which is in line with the main purpose of the portfolio. This element relates to the credibility strategy of triangulation.

Dependability can be reached by establishing an audit trail and by the use of external auditors. Both strategies were used to monitor our assessment procedure. The audit trail consisted of comprehensive documentation of the different steps of the assessment process: a formal assessment plan approved by the examination board; portfolio and assessment guidelines; overviews of the results per phase, and written assessment forms per student. Quality assurance procedures were set up. The internal quality procedure enables a student to appeal to the university Board of Appeal for Examinations against the outcome of the assessment. The external quality procedure entails regular audit by the Dutch organisation for educational auditing and accreditation. This relates to the dependability strategy of audit.

The case study showed that all these elements contributed to the credibility and dependability of portfolio assessment. We are convinced that this assessment process has considerably more credibility and dependability than procedures aimed at high interrater reliability, particularly if such procedures necessitate standardisation and rigid structuring with concomitant impairment of validity. In the hypothetical case of a legal procedure, we would be able to build our defence on the evidence that after a careful assessment procedure a committee of experts had reached consensus on the final decision on a student's portfolio. A better defence is hard to imagine. Further support for this assessment procedure might be sought by means of a study involving different assessment committees.

Although the mentoring process was resource-intensive, most of the mentors' time was spent on mentoring and formative feedback and only a minor portion on formal assessment. During an informal debriefing the mentors indicated that the judgement procedure had not burdened them disproportionately.

To some extent, the portfolio assessment procedure described in this paper resembles the procedure suggested by Friedman Ben David *et al.,* who proposed 2 independent ratings followed by a final consensus procedure, culminating in an overall judgement and agreement between the 2 raters.[1] Our procedure is broader and includes information collected from the very start of the portfolio compilation process, the mentor's and the student's input, as well as a final full committee consensus decision.

At the heart of the approach used in our case study is the recommendation by McCullan *et al.* in their review of portfolio studies, that, for portfolio assessment, criteria from qualitative research might be more appropriate than criteria from quantitative research, like reliability.[3] Instead of looking at consistency across repeated assessments (a quantitative psychometric approach), we added information to the judgement process until saturation of information was reached (a qualitative approach).[20] This does not mean that psychometric aspects were ignored. In fact, the concept of psychometrics – particularly in relation to sequentially increasing the number of examiners – was used, but it was not applied in a classic test theoretical sense.

Naturally, some arbitrary decisions were unavoidable. It is not unlikely that adaptations of the procedure will be proposed, depending on our cumulative experiences and evaluations. However, the essence of our argument is that we have tried to demonstrate that reliability, as viewed from the purely psychometric perspective, is too limited a criterion to be applied to qualitative assessment and that reliability can be built into an assessment procedure by implementing various safeguards and by collecting more information only when necessary. We believe that this represents an important step forward in our endeavours to incorporate more qualitative and subjective features into competence assessment.

*Contributors:* All authors contributed ideas to the paper and commented on all drafts of the paper. ED designed and is co-ordinator of the portfolio programme in Years 1–4 of the Maastricht undergraduate curriculum.
*Acknowledgement:* we would like to thank Mereke Gorsira for her help in improving the English language of the paper.
*Funding:* none.
*Conflicts of interest:* none.
*Ethical approval:* not applicable.

## REFERENCES

1 Friedman Ben David M, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ. AMEE Medical Education Guide, 24. Portfolios as a method of student assessment. *Med Teacher* 2001;**23** (6):535–51.
2 Webb C, Gray M, Jasper M, McCullan M, Scholes J. Models of portfolios. *Med Educ* 2002;**36**:879–98.
3 McCullan M, Endacott R, Gray MA *et al.* Portfolios and assessment of competence: a review of the literature. *J Adv Nurs* 2003;**41** (3):283–94.
4 Snadden D. Portfolios – attempting to measure the unmeasurable? *Med Educ* 1999;**33**:478–9.
5 Baume D, Yorke M. The reliability of assessment by portfolio on a course to develop and accredit teachers in higher education. *Studies Higher Educ* 2002;**27** (1):7–25.
6 Pitts J, Coles C, Thomas P. Educational portfolios in the assessment of general practice trainers: reliability of assessors. *Med Educ* 1999;**33**:515–20.
7 Roberts C, Newble DI, O'Rourke A. Portfolio-based assessments in medical education: are they valid and reliable for summative purposes? *Med Educ* 2002;**36**:899–900.
8 Koretz D, Stecher B, Klein S, McCaffrey D. The Vermont Portfolio Assessment Program: findings and implications. *Educational Measurement: Issues Pract* 1994;**13** (3):5–16.
9 van der Vleuten CPM, van Luijk SJ, Ballegooijen AMJ, Swanson DB. Training and experience of medical examiners. *Med Educ* 1989;**22**:290–6.
10 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.
11 Huot B. Beyond the classroom: using portfolios to assess writing. In: Black L, Daiker DA, Sommers J, Stygall G, eds. *New Directions in Portfolio Assessment. Reflection, Practice, Critical Theory and Large-scale Scoring.* Portsmouth, New Hampshire: Cook Publishers 1994;325–33.
12 van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;**25**:110–8.
13 Lincoln YS, Guba EA, eds. *Naturalistic Inquiry.* Beverly Hills, California: Sage 1985.
14 Denzin NK, Lincoln YS, eds. *Handbook of Qualitative Research.* 2nd edn. Thousand Oaks, California: Sage 2000.
15 Webb CER, Gray M, Jasper M, McMullan M, Scholes J. Evaluating portfolio assessment systems: what are the appropriate criteria? *Nurse Educ Today* 2003;**23**:600–9.
16 Norcini JJ, Shea JA. The credibility and comparability of standards. *Appl Measurement Educ* 1997;**10** (1):39–59.
17 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.
18 Driessen EW, van der Vleuten CPM. Matching student assessment to problem-based learning: lessons from experience in a law faculty. *Studies Cont Educ* 2000;**22** (2):235–48.
19 Zeichner K, Wray S. The teaching portfolio in US teacher education programmes: what we know and what we need to know. *Teaching Teacher Educ* 2001;**17** (5):613–21.
20 Strauss AL. *Qualitative Analysis for Social Scientists.* New York: Cambridge University Press 1987.

*Received 22 September 2003; editorial comments to authors 18 November 2003, 5 March 2004; accepted for publication 15 April 2004*