Assessment & Evaluation in Higher Education, Vol. 28, No. 4, 2003

Carfax Publishing
Taylor & Francis Group

Developing Performance Standards for Teacher Assessment by Policy Capturing

MARIEKE F. VAN DER SCHAAF, KAREL M. STOKKING, *Department of Educational Sciences, Utrecht University, Utrecht, The Netherlands*

NICO VERLOOP, *ICLON Graduate School of Education, Leiden University, Leiden, The Netherlands*

ABSTRACT *There is a need for assessment of teachers' competencies fostered by a growing attention given to accountability and quality improvement. Important questions are how good the demonstrated competencies of teachers should be for a satisfying assessment and how the different competencies should be weighted. Using a policy capturing method, in two rounds, nine stakeholders developed performance standards (or cut-off scores) for teacher assessment on eight criteria (or content standards) that resulted from an earlier study. Between the rounds, the panellists held a structured group discussion. Policy capturing proved to be a clear and useful method generating consistent judgements that can be described according to both a compensatory model and a conjunctive model. From the first to the second round, the consistency increased. However, while the panellists agreed to a substantial degree on the performance standards, they disagreed on the weights to be assigned to the criteria.*

Introduction

There is a demand for accountability and quality improvement in the teaching profession. Teachers are expected to display the competencies (knowledge, skills and attitudes) needed to perform their tasks and to engage in continuous professional development. Assessment is used as a tool to ascertain whether teachers satisfy the required competencies and to formulate guidelines for professional development. In the first case, assessment is used summatively in order to account for the teacher's quality, with possible consequences such as certification and merit pay. In the second case, assessment has a formative goal and produces information that can be used for planning activities directed at further professional development.

ISSN 0260-2938 print; ISSN 1469-297X online/03/040395-16 © 2003 Taylor & Francis Ltd
DOI: 10.1080/0260293032000066227

It is common to describe competencies in terms of criteria (or content standards) and performance standards. The criteria describe the dimensions on which teachers should be assessed and performance standards are the cut-off scores that indicate the minimum levels for a sufficient result. In this article, we focus on the development of performance standards for the purpose of summative assessment of experienced teachers.

Over the last decade, teacher competencies have been increasingly assessed by performance assessment, accompanied by a change in instruments from behavioural observations and knowledge tests to simulations and portfolios. The latter instruments recognise that teaching is a complex activity and that teacher behaviour is inextricably bound up with the teacher's cognition and with the situation in which the teaching takes place. In performance assessment, the judgement is preferably not dichotomous (in terms of right or wrong) but uses multiple cut-off scores (e.g. basic, competent, advanced). Additionally, it is usual to judge on several criteria, which are possibly differentially weighted.

In the last 10 years, different methods have been developed for generating performance standards to be used in performance assessment (Berk, 1996). Many methods are criterion referenced, in which the candidate's score is directly compared with a performance standard. The performance is judged satisfactory if it surpasses this standard. Most of these criterion-referenced methods are based on stakeholders' judgements, in which stakeholders are asked to assess the minimum score necessary to satisfy the criteria.

Some methods using stakeholders are based on policy capturing (see for example Jaeger, 1995; Putnam *et al.*, 1995). Policy capturing has been used in various fields to generate insight into the judges' cognitive processes (Cooksey, 1996). This method is particularly interesting as it is not only directed towards the question of how well teachers should score but also towards the weighting of the criteria and the specific judgemental model. To answer these questions, policy capturing uses multiple regression procedures to produce a best fitting equation or policy. Stakeholders judge a great number of teacher profiles, being configurations of scores on criteria. The judgements are analysed by multiple regression analysis, yielding a policy that indicates the performance standards and the criteria weights. This procedure can be used both for the whole panel of stakeholders (panel policy) and the individual panel members (personal policies).

The merits and shortcomings of using policy capturing for developing performance standards are still largely unknown and the implementation meets some obstacles (Hambleton, 1997), of which we mention two. First, the method is a difficult one for the participating stakeholders, even after training. This is problematical because the validity of the performance standards depends to a high degree on the stakeholders' competence to implement the method. Second, the method aims to reach consensus between the panellists. It is likely that panellists' judgements are based on different assumptions and perspectives, which probably result in different preferences for performance standards. To reach consensus all panellists should agree with the performance standards that best meet the different underlying assumptions and perspectives. Whereas the resulting performance standards will be based on a mix of different underlying visions, the precise meaning of the standards will become less clear.

In this study, we develop and test a policy capturing method that aims to overcome these problems. The study is part of a larger research project on the summative judgement of experienced social sciences teachers' competencies (economics, geography, history) in pre-university education in The Netherlands (ages 16–18), comparable with undergraduate college level education in, for example, the American system. The teachers are expected to promote the development of their students research skills by having them work independently and in groups on assignments that result in a report or presentation.

Developing Performance Standards

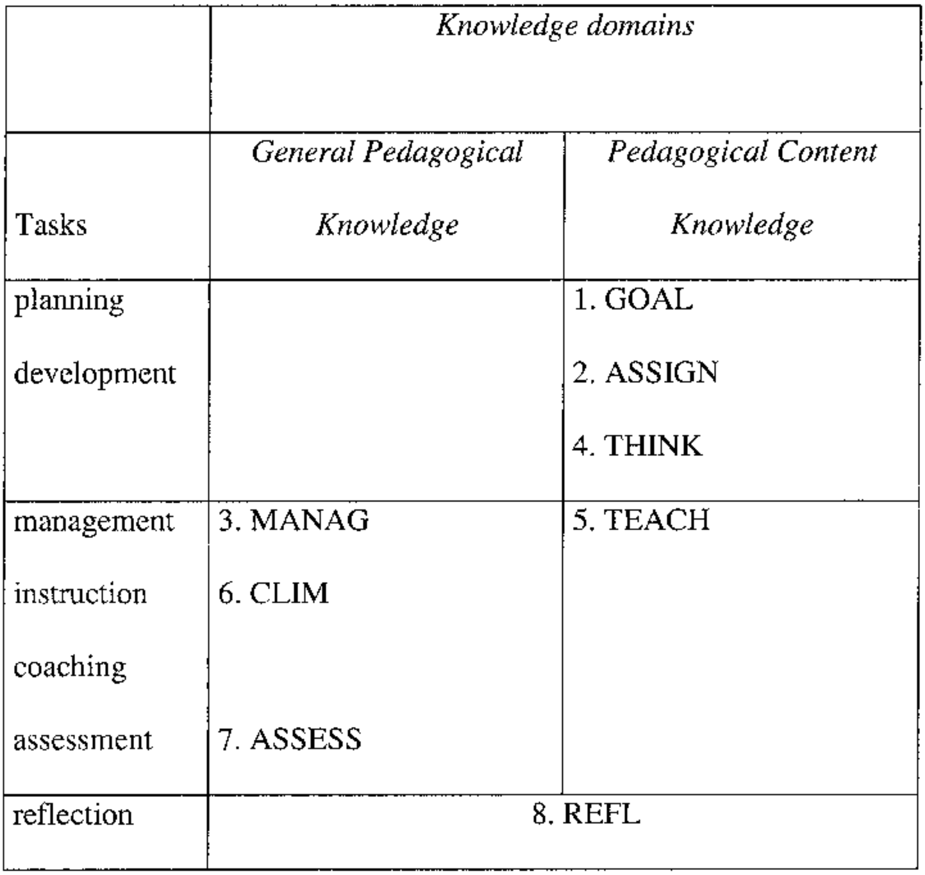
Criteria

Competence is a generic term for the knowledge, skills and attitudes required for adequate functioning in a profession. Competence can be divided into the (sub)competencies needed to satisfactorily perform a set of tasks. Competencies can be made visible in actions and be influenced by the context (Gonczi, 1993) and are seen as combinations of behaviour and the underlying intentions and motives. Following this definition in an earlier Delphi study, we developed criteria that became the input for this study (van der Schaaf *et al.*, 2003). The criteria have been formulated both in terms of tasks and knowledge domains. The combination of both dimensions generates a matrix. One axis describes the tasks that teachers have to fulfil when teaching. The other axis describes the knowledge domains that teachers rely on when they fulfil their tasks. The cells refer to the criteria that teachers have to meet when teaching research skills to students (see Figure 1).

The criteria are as follows. (1) Selecting year long goals for students (GOAL): teacher's long-term goals for developing students' research skills and his or her approach to reach these goals. (2) Choosing an appropriate assignment (ASSIGN): the goals, content and form of the assignment. (3) Preparing and managing students working on their assignments (MANAG): the preparation, communication and management of facilities (time, rooms, sources, media, etc.) that students need for fulfilling the assignment. (4) Previously thinking of and selecting teaching strategies that support the development of research skills of students (THINK): the choice of and argumentation for the use of teaching strategies that meet students' knowledge, abilities and experience. (5) Teaching students research skills (TEACH): the use of the teaching strategies chosen as described in criterion 4. (6) Creating a positive pedagogical climate (CLIM): the provision of a safe, respectful and stimulating learning environment to students. (7) Adequately assessing research skills of students (ASSESS): teachers' argumentation for and the clarity and comprehensibility of the assessment approach used (criteria, scoring and norms), the applicability of the approach to the assignment and the way the teacher handles the assessment and communicates the results to the students. (8) Reflecting on the programme and on actions in teaching students research skills (REFL): the extent to which the teacher is aware of strong and weak points in his or her teaching and his or her suggestions for improvement.

Judgemental Model

High stake assessment, such as for certification, selection or merit pay, must result in a single outcome. For that purpose, the scores on the criteria have to be combined into a final judgement. Two models that are frequently used for combining scores on criteria are the compensatory model and the conjunctive model. In the compensatory model, the candidate should achieve a defined total score to pass the assessment, allowing high scores on certain criteria to compensate for low scores on other criteria (Mehrens, 1990).



	Knowledge domains	
Tasks	*General Pedagogical Knowledge*	*Pedagogical Content Knowledge*
planning development		1. GOAL 2. ASSIGN 4. THINK
management instruction coaching assessment	3. MANAG 6. CLIM 7. ASSESS	5. TEACH
reflection	8. REFL	

FIG. 1. Criteria in function of tasks and knowledge domains.

The conjunctive model requires that the candidate attains a certain minimum score on each criterion.

Choosing between these models one should consider the following: (i) the perception of teaching as a profession. Is teaching a profession in which it is natural to have both strong and weak sides or should a teacher pass minimum requirements for each relevant criterion to guarantee a competent practice?; (ii) the supposed linkages between the criteria. If the criteria are strongly related, it is reasonable to let them compensate for one another; if not, a conjunctive model is a sensible choice; (iii) the avoidance of incorrect judgements. With a compensatory model, positive and negative deviations cancel each other. With a conjunctive model, only one good or bad score, as a result of good luck or bad luck, could produce an incorrect overall judgement.

We prefer the compensatory model. It is realistic to assume that teachers have weak sides and that they still have scope to develop further. In addition, we focus on high stake performance assessment, attempting to avoid incorrect judgements as much as possible.

Professional Development

Generally, performance standards are based on an explicit model of the processes behind the competencies or the way these competencies develop. Often, development is placed on a continuum from less to more professional. Aiming at teaching performance standards, this approach is not wholly adequate, because development can be an irregular process with sudden improvements or unexpected environmental influences (Huberman, 1995; Day, 1999). In addition, the course of development can differ between individuals and between contexts (Eraut, 1994). Performance standards can also be based on minimally required or maximally feasible mastery levels (Wheeler & Haertel, 1993), as we do in our study.

Quality Requirements for Performance Standards

Kane (1994) describes three types of evidence for demonstrating the appropriateness of performance standards which are developed judgementally, based on stakeholders' assessments: (i) validity checks based on external criteria consisting of comparisons of the performance standards with external information about the desired teacher competencies. As our study is directed towards a rather new element in the curriculum and external information is not yet available, we restrict ourselves to the other two types of evidence; (ii) validity checks based on internal criteria, concerning the consistency of the judgements in the standard setting procedure. The consistency indicates the reliability of the procedure used. We deliberately use the word 'indicates', because variation in panellists' judgements does not necessarily mean that the procedure as such is not reliable, since the panellists could interpret the performance standards differently; and (iii) procedural evidence, consisting of demonstrating the soundness of the chosen performance standard setting procedure and its use, including the selection of the panellists involved in the process.

Method

Sample of Stakeholders

A panel of nine stakeholders was selected. This number is assumed to be sufficient for setting reproducible standards (Berk, 1996). Seven stakeholders also participated in an earlier study in which criteria for teaching students research skills were developed (van der Schaaf *et al.*, 2003). Hence, these panellists were familiar with the criteria. The additional two panellists were thoroughly briefed.

The panel consisted of three school principals (one of them formerly a history teacher and teacher trainer), two experienced geography teachers who were also teacher trainers, two experienced history teachers, one experienced teacher in both geography and history and an experienced teacher in social science subjects who is also a teacher trainer in economics. The teacher trainers were working at teacher training institutes at or allied with several universities. All panellists were familiar with the objectives and approaches of teaching research skills to students aged 16–18. They supported the basic assumptions and goals of our study and received financial compensation. None of the panellists had ever participated in a policy capturing study before.

Task

The panellists' task involved judging fictitious teacher profiles. Each profile consisted of

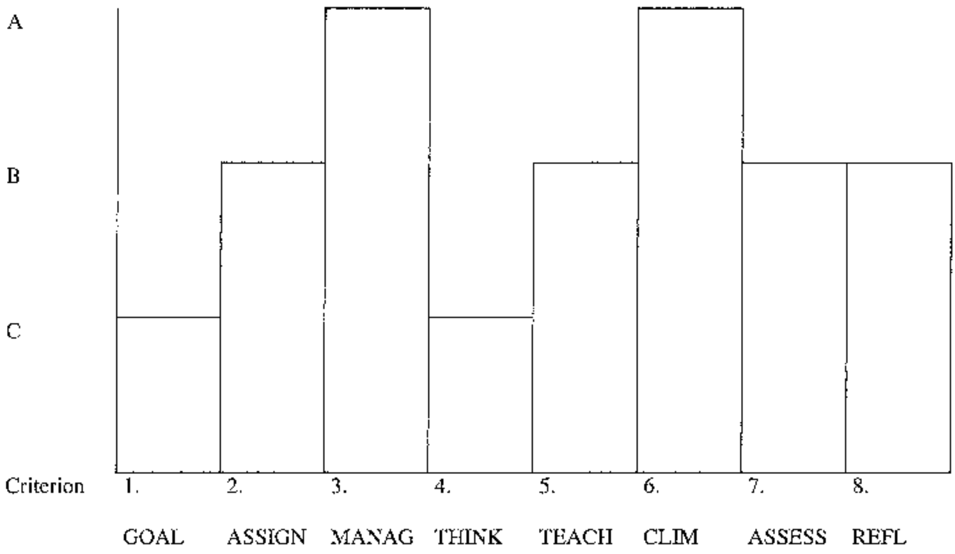


FIG. 2. Example of a profile. A, the teacher satisfies the criterion completely or to a strong degree; B, the teacher satisfies the criterion to a reasonable or to some degree; C, the teacher satisfies the criterion to a small degree or not at all.

a configuration of fictitious scores of one teacher on the eight criteria shown in Figure 1. That number of criteria is close to the maximum number that judges can handle in a policy capturing procedure (Cooksey, 1996). Each criterion met one out of three anchor points, namely the teacher satisfies the criterion: A, completely or to a strong degree; B, to a reasonable or to some degree; C, to a small degree or not at all. The profiles varied in their scores on the criteria. See Figure 2 for an example.

In our study, we focused on portfolio assessment, a portfolio being taken as a collection of material about the teacher's work, in this case concerning a research assignment the teacher has given to the students. The material was made up of a self-description, the assignment, interviews, videos of teaching research skills to students, evaluations by students and judgements of students' work by the teacher. Before starting their task, the panellists studied an example portfolio. They also studied a manual in which all eight criteria, all anchor points for these eight criteria and the portfolio material had been carefully and comprehensively described.

In a first round, the panellists judged 64 profiles. They judged 40 profiles in a second round. In the second round, 13 profiles from the first round were repeated to investigate whether the panellists' judgements were consistent (fitting the recommended number for replicated profiles) (Cooksey, 1996). The panellists were not informed of this. In order to minimise the effects of memory, we mixed the 13 replicated profiles with the 27 new profiles randomly.

Sample of Profiles

Ideally, the sample of profiles represents the target domain and is drawn from empirical cases. However, in our study this was not feasible because teachers were still inexperienced in teaching students research skills. Therefore, we used a sample of fictitious profiles. With 8 criteria and 3 anchor points 3^8 or 6561 profiles can be constructed. From these, using SPSS, we drew random samples of 64 and 40 profiles to be used in the first and second rounds, respectively (mean scores 2.02 and 1.96; standard deviations 0.29 and 0.30; minima 1.38 and 1.25; maxima 2.63 and 2.50). The ratios between the number of profiles and the number of criteria were 8 : 1 and 5 : 1, respectively, and match with the recommended minima (Cooksey, 1996).

Procedure

As policy capturing is a difficult task and it is unclear how panellists having different perspectives could reach consensus, we used several means to support the process. Firstly, the panellists were given feedback on their judgements to increase their understanding of the judgemental process (Van de Pligt, 1997). Secondly, the panellists were given opportunities to reconsider their performance standards (Cross *et al.*, 1984). Thirdly, the panellists discussed their judgements (Fitzpatrick, 1989). Fourthly, procedures were included to prevent negative effects of group discussion as much as possible. The procedure consisted of 6 steps.

1. Training (4 hours plenary). The panellists studied a manual describing the objectives, planning and procedures of policy capturing. They then followed a training focused on the panel's aim, task and procedure, including a discussion about the criteria, the anchor points, and the portfolio. The training goal was to let them practice judging the profiles, without standardising their way of judging. The judged profiles were discussed, including the extent to which the panellists agreed and the reasons for their judgements.

2. Judgements, first round (3 hours individually). The panellists received a booklet containing 64 profiles. It was explained that the profiles were fictitious. The panellists were asked to judge the profiles on the 3 anchor points. The panellists judged the profiles individually at home.

3. Studying feedback (1 hour individually). The panellists received written feedback, consisting of a graph illustrating the panel policy and the personal policies resulting from the multiple regression analysis.

4. Discussion (4 hours plenary). The feedback (step 3) was the input for a meeting in which the panellists discussed the underlying arguments for the different policies and tried to reach consensus.

In group discussion, participants can be influenced by social comparison and by information from others. During the meeting, we tried to minimise the effects of social comparison (which would bring about unwanted effects, such as groupthink, polarisation and domination by a few panel members) and to promote the effects of information (which would increase the quality of the discussion). We used the nominal group technique (Delbecq *et al.*, 1975), consisting of a structured discussion in which the panellists alternate between working individually and in plenary sessions. During the plenary sessions they clarify their arguments and try to convince the other participants of their judgements.

Public commitment to their personal policies could make panellists resistant to changing their opinion (Fitzpatrick, 1989). To control for this effect, we divided the panel during step 4 (Discussion) into two subgroups: five panellists who participated in the group discussion (panellists 1, 2 and 6–8) and four panellists who had an individual discussion with a researcher (the first author) (panellists 3–5 and 9). The group discussion and the discussions with the researcher, which were both video-recorded and fully transcribed, followed the phasing of the nominal group technique: (i) after recognising their personal policies, the panellists considered how to substantiate them; (ii) each panellist presented and clarified his or her personal policy (to the other participants or to the researcher); (iii) the panellists concluded their presentation in a few statements summarising the main points. This resulted in a list of 26 statements; (iv) the panellists that participated in the group discussion examined whether they could reach agreement on a panel policy. In addition, all panellists judged the 26 statements (the result of step c) on a 5-point scale (from greatly agree to greatly disagree).

5. Judgements, second round (2 hours individually). As panellists reach better results if they have the opportunity to change their earlier policies (Jaeger, 1990), they were asked again to judge 40 fictitious profiles. They received the same instructions and were asked the same questions as in the first round.

6. Feedback. After the second round of judgements the panellists received feedback on their personal policy and the panel policy.

Data Analysis

The panellists judged fictitious profiles of teachers, each profile consisting of scores on eight criteria. In the multiple regression model the profiles' scores on the criteria are the predictor variables and the panellists' judgements of the profiles are the criterion variables. For multiple regression analysis to be valid, assumptions of normality of the residuals, homoscedasticity, linearity and no collineation have to be fulfilled. We checked these by: making a histogram of the standardised residuals from the judgements of the profiles, making a plot of the standardised residuals against the predicted judgements and calculating the tolerance of the criteria and the correlations between the criteria.

Our next questions were: (i) whether teachers' competencies can be best judged according to a compensatory or a conjunctive model; (ii) what weights the criteria should have in the judgement; (iii) what the minimum overall profile score should be for teachers to be judged as 'satisfactory'. To answer these questions we executed multiple linear regression analyses according to a compensatory model and multiple regression analyses applying a natural logarithm transformation according to a conjunctive model (Jaeger, 1995; Cooksey, 1996). To indicate the quality of the models, we used the proportion explained variance (r^2). The closer r^2 is to 1.0, the better the model describes the judgements.

To reveal possible differences between personal policies, we depicted them in a diagram and computed the standardised β weights of the criteria in the personal policies.

In addition, we checked the consistency of the panellists' judgements by computing the correlations between the judgements of the 13 replicated profiles between the first and second round. Panellists possibly judged more consistently in the second round as a result of the group discussion. To control for this, we used the Mann–Whitney test test (2-tailed) to test for differences in the consistency between the panellists who participated in the two subgroups.

Further, we analysed the panellists' scores on the 26 statements summarising their main assumptions and perspectives. We searched for a relation between the panellists' perspectives and their personal judgemental policies by computing the partial correlations between the scores on the statements and the standardised β values of the personal policies.

Finally, the panellists evaluated the policy capturing method on the following aspects: the difficulty of the procedure, the recognisability of their personal policy and the role achieved by the feedback.

Results

Assumption Checks

The assumptions of normality, homoscedasticity and linearity were sufficiently met. The criteria (the predictors in the model) were barely collineated: the tolerance was high (between 0.80 and 0.95 with one outlier of 0.75) and the Pearson correlations were low (between $r = -0.34$ and $r = 0.20$, mostly around $r = 0.00$).

Judgemental Model, Criteria Weights and Performance Standards

Firstly, the panel policy and the personal policies for both rounds were analysed. Table 1 shows that the conjunctive model predicts panellists' judgements slightly better than the compensatory model. However, in our view the difference between both models is not large enough to relinquish the advantages of the compensatory model. The higher r^2 in the second round indicates that the model better represents the judgements in the second round than in the first round.

Secondly, we wanted to know what weights the criteria should be given. Table 2 contains the results of the analyses, using a compensatory model. The β values show that the panel gives criteria 5 (TEACH) and 6 (CLIM) more weight than the other criteria.

Thirdly, the panel policy was calculated in order to determine what the minimum total score should be for teachers to be judged as 'satisfactory'. Every profile consisted of 8 criteria with (variable) scores on anchor point A, B or C. The profiles were judged by the panellists as 'good', 'satisfactory' or 'not satisfactory'. To calculate the panel policy we attached the scores 3, 2 and 1 to the anchor points A, B and C and also 3, 2 and 1 to the judgements 'good', 'satisfactory' and 'not satisfactory'. So, both the average profile score (the average of the fictive scores on the 8 criteria) and the judgements can

TABLE 1. Adequacy of the panel policy and personal policies in terms of r^2

	Round 1		Round 2	
Panellist	Compensatory r^2	Conjunctive r^2	Compensatory r^2	Conjunctive r^2
1	0.70	0.71	0.62	0.67
2	0.43	0.45	0.60	0.71
3	0.40	0.44	0.74	0.77
4	0.51	0.53	0.74	0.79
5	0.45	0.55	0.78	0.80
6	0.62	0.73	0.55	0.68
7	0.71	0.73	0.70	0.76
8	0.80	0.74	0.88	0.89
9	0.55	0.59	0.80	0.81
Panel (total)	0.37	0.40	0.54	0.58

TABLE 2. Results of the multiple regression analyses in rounds 1 and 2, compensatory model

	Round 1 ($n = 575$)			Round 2 ($n = 360$)		
	b	Standardised β	t ($P < 0.001$)	b	Standardised β	t ($P < 0.001$)
1 GOAL	0.13	0.16	4.40	0.16	0.19	4.68
2 ASSIGN	0.15	0.19	5.11	0.23	0.27	6.98
3 MANAG	0.15	0.18	5.19	0.06	0.07	1.83[a]
4 THINK	0.12	0.17	4.45	0.15	0.17	4.10
5 TEACH	0.25	0.33	9.55	0.33	0.38	9.78
6 CLIM	0.21	0.27	7.66	0.33	0.35	8.66
7 ASSESS	0.15	0.19	5.34	0.19	0.21	5.10
8 REFL	0.10	0.13	3.72	0.14	0.15	3.85
Constant	− 0.73 (se 0.16)		− 4.63	− 1.35 (se 0.18)		− 7.64
Round 1	$r = 0.61$	F (df) = 41.49 (566) ($P < 0.001$)		$r^2 = 0.37$	Adjusted $r^2 = 0.36$	
Round 2	$r = 0.73$	F (df) = 50.98 (351) ($P < 0.001$)		$r^2 = 0.54$	Adjusted $r^2 = 0.53$	

[a] $P = 0.69$.

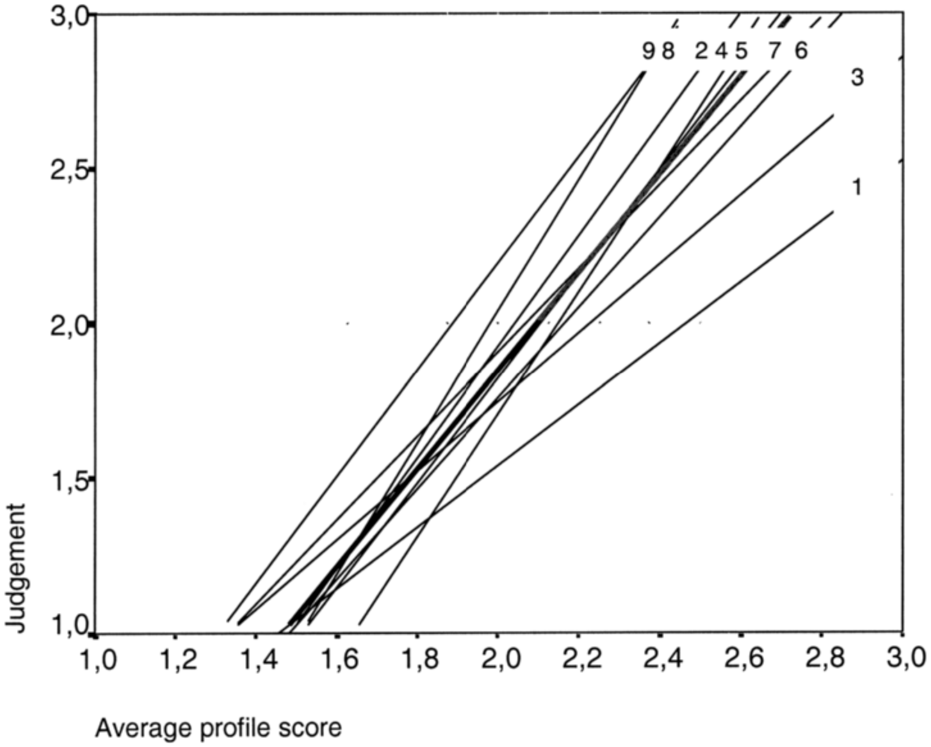


FIG. 3. Judgements of mean profile scores by panellists in round 2, compensatory model ($r^2 = 0.54$). The numbers near the policy lines refer to the panellists. The bold line refers to the panel policy. Judgement (y-axis): 1, not satisfactory; 2, satisfactory; 3, good. Algorithm of the panel policy: 0.16 × GOAL + 0.23 × ASSIGN + 0.06 × MANAG + 0.15 × THINK + 0.33 × TEACH + 0.33 × CLIM + 0.19 × ASSESS + 0.14 × REFL − 1.35.

meet scores between 1 and 3. In Figure 3, the x-axis represents the average profile scores and the y-axis refers to the judgements given. The bold line indicates the panel policy (round 2, compensatory model). The thin lines indicate the personal policies. Figure 3 shows that some of the panellists judge more leniently than others. However, most of the thin lines are close together, so the panellists as a group agreed to a great extent on the performance standards. The results in the first round (not shown here) were very comparable. In both rounds, teachers needed minimally an average profile score between 2.1 and 2.2 to be judged as 'satisfactory' (i.e. to get a mean panel judgement of 2).

In our study 'satisfactory' means that a teacher's portfolio is sufficient and meets the panel's perception of a competent teacher to a moderate to high degree. A 'satisfactory' total score can have several consequences, like certification or merit pay in the case of summative assessment or insight into strong and weak points that need further development in the case of formative assessment.

Although most panellists were close together on the cut-off point, their opinions differed as to the weights to be attached to the criteria. Table 3 gives a survey per round of the weights the panellists attached to the criteria, based on the compensatory model.

Most of the panellists attached comparable weights to the criteria in the second round as they did in the first round. The panellists' answers were not related to their professional position (e.g. teacher, principal, teacher educator) or to their discipline (economics, geography, history) or to other characteristics.

Most of the correlations (Pearson) between the judgements of the 13 replicated profiles in the first and second round were $r > 0.75$ ($P < 0.05$) (with two outliers of $r = 0.55$ and $\mathrm{r} = 0.63$), indicating consistency of the panellists' judgements. According to the Mann–Whitney test (2-tailed), there was no significant difference between the panellists who participated in the group discussion ($n = 5$) and those who had an individual conversation with the researcher ($n = 4$).

The Panel's Underlying Assumptions and Perspectives

Each panellist summarised the underlying assumptions and perspectives of his personal policy in a few statements. The resulting 26 statements were scored by the panellists on a 5 point scale (from 1 = I greatly agree to 5 = I greatly disagree). Exploratory factor analysis followed by item analyses resulted in three sufficiently reliable Likert scales, each consisting of five statements. We illustrate each scale with an example of a statement: (i) the extent to which the panellists connect the criteria to each other, e.g. 'all criteria are connected, without the others they could not exist' (Cronbach's α 0.85; mean 2.42; minimum 1.75; maximum 3.44; variance 0.54); (ii) the extent to which the panellists want to reduce the judgemental model to one core, e.g. 'to sharpen the judgemental model the number of criteria should be reduced' (Cronbach's α 0.84; mean 1.95; minimum 1.67; maximum 2.44; variance 0.13); (iii) the extent to which the panellists are directed at the learning environment of students, e.g. 'the choice of a research assignment is important because a good assignment is a condition for students' research skills to develop' (Cronbach's α 0.74; mean 2.22; minimum 1.44; maximum 3.33; variance 0.47).

The panellists' answers were not related to their professional position or to their discipline nor to their personal judgemental policies. In some cases panellists had different reasons for the same judgements, while in other cases panellists gave different judgements for the same reasons. This observation was confirmed by the correlations between the scores on the scales with statements and the standardised β values expressing the personal policies in the second round. Only 2 out of 24 correlations (3 scales × 8 criteria) were significant.

The Panel's Evaluation of the Policy Capturing Procedure

The panellists thought judging the profiles not to be very difficult nor easy. On a 5 point scale (1 = easy; 5 = difficult) the mean score was 2.9 in the first round and 2.6 in the second round. Profiles with unequivocal positive or negative scores on important criteria were easy to judge, whereas it was difficult to judge profiles with both positive and negative scores on important criteria. The panellists judged the latter in some cases not to be realistic or to be rarely found in practice. Secondly, almost all panellists recognised themselves in their personal policy. Thirdly, for all panellists the feedback given to them confirmed what they had previously thought to be important in judging teachers.

Conclusions

In this study, the focus was on the development of performance standards for high stake teacher assessment purposes. During the 1990s several performance standard setting

TABLE 3. Significant ($P \leq 0.05$) standardized βvalues per criterion per panellist

Panellist	Round	1. GOAL	2. ASSIGN	3. MANAG	4. THINK	5. TEACH	6. CLIM	7. ASSESS	8. REFL
1	1	0.43	0.21	[a]	0.24	0.39	0.36	0.24	[a]
	2	0.27	[a]	[a]	[a]	0.44	0.41	[a]	[a]
2	1	[a]	0.22	0.22	[a]	0.36	0.37	[a]	0.23
	2	[a]	0.21	[a]	[a]	0.39	0.52	[a]	0.33
3	1	0.24	[a]	[a]	0.26	[a]	0.33	0.38	[a]
	2	[a]	0.40	[a]	[a]	[a]	0.44	0.43	[a]
4	1	0.28	[a]	0.30	[a]	0.30	0.30	[a]	0.30
	2	0.34	[a]	[a]	0.35	0.31	0.36	[a]	0.42
5	1	0.20	0.21	0.23	[a]	0.37	0.26	[a]	0.28
	2	0.20	0.32	[a]	0.20	0.50	0.37	0.31	[a]
6	1	[a]	0.67	[a]	[a]	[a]	[a]	0.41	[a]
	2	[a]	0.51	[a]	[a]	[a]	[a]	0.30	[a]
7	1	0.25	[a]	0.19	0.22	0.51	0.54	[a]	[a]
	2	0.22	[a]	[a]	[a]	0.53	0.51	[a]	[a]
8	1	0.19	0.27	0.34	0.25	0.38	0.30	0.36	0.19
	2	0.31	0.42	0.22	0.25	0.23	0.31	0.39	[a]
9	1	0.30	[a]	0.18	0.22	0.61	[a]	0.22	[a]
	2	[a]	0.24	0.22	0.37	0.75	[a]	0.19	[a]

[a] Not significant ($P > 0.05$)

methods were developed that can be used for performance assessment. Those methods based on policy capturing are particularly interesting because they can be used for both the development of performance standards and to gain insight into the judgemental process (compensatory or conjunctive) as well as into the weights to be attached to the criteria (Jaeger, 1995). The disadvantages of this method are its difficulty for the judges and the fact that it is not clear how to foster consensus (which is seen as an indicator of success). In this study we attempted to counter these difficulties by: a careful selection of panellists representing different opinions and having sufficient knowledge and motivation for their task; construction of the judgemental task from eight criteria about which panellists in an earlier study had reached consensus; a training process; interim feedback; a group discussion using the nominal group technique; judgement of profiles in two rounds.

To indicate its usefulness, we now summarise the main merits and shortcomings of the policy capturing procedure, as utilised in our study.

Firstly, the procedure produced significant and consistent regression models. The panellists' judgements in the second round were more consistent than in the first round. It is not precisely clear which intervention or combination of interventions contributed to this growth in consistency.

Secondly, we investigated whether we should follow a compensatory or conjunctive model for combining subscores into a final judgement. Whereas it is realistic to assume teachers have weak points that they still have to develop further and it is important to avoid incorrect judgements caused by errors, we initially preferred a compensatory model. The results show that both models represent the panellists' judgements nearly equally well. We conclude that there is no reason to abandon our preference for a compensatory model.

Thirdly, the panellists agreed rather strongly on the performance standards accompanying the eight criteria. In both rounds, the average profile score needed for the judgement 'satisfactory' turned out to lie between 2.1 and 2.2. Hence, according to the algorithm of the compensatory model, the cut-off profile score is 2.2. This outcome can be used for portfolios that are assessed per criterion on a 3-point scale (between 1 and 3).

Fourthly, the panellists did not agree on the weights to be attached to the criteria in the final judgement. This result differs from that of another study on developing performance standards (Hambleton & Plake, 1995) in which a judgement procedure in two rounds with interim discussion resulted in high agreement on the panel policy. This discrepancy might be explained in several ways. The nominal group technique, in which panellists publicly present their personal policies, could discourage individuals from changing their initial response (Fitzpatrick, 1989). However, the controls conducted did not show any evidence for this. Next, it is possible that the feedback given has been perceived after all as supporting the initial personal ideas. Additionally, our study might have exerted less pressure to revise the initial policy because the panellists knew they were participating in a research project and not in an authentic summative assessment context. Also, the time span per round was possibly too short for the panellists to revise their opinions. Policy capturing could be more effective if it is undertaken by panellists working in the same context, sharing the same frame of reference and over a longer timespan.

Fifthly, our stakeholders perceived the eight criteria as being interwoven and would welcome a reduction in their number. This is not an unknown phenomenon. For example, the academic literature on assessment centres reveals that during the assess ment, judges merge criteria, prefer certain criteria or have difficulty distinguishing criteria (Arthur *et al.*, 2000). The interventions in our study to familiarise the panellists with the criteria, such as the training and the manual, were apparently not sufficient to prevent this.

Sixthly, in our study policy capturing did not prove to be a difficult procedure for the participants. Profiles with consistently high or low scores on important criteria were relatively easy to judge, whereas profiles with 'inconsistent' scores were perceived to be more difficult, especially if the panellists thought the profiles were unrealistic. Empirical profiles might prevent this problem.

We conclude that the policy capturing method, as used in our study, is a useful method for stakeholders to develop performance standards. The method is feasible for the panellists and leads to consistent performance standards. The procedure followed brings about a learning effect, resulting in more consistent judgements.

This research should be seen in the context of a growing interest in teacher assessment and the need for suitable methods to develop performance standards. We wanted to contribute by developing and testing a policy capturing method. The need for new methods and for research on this topic will remain for some time. It is important to fulfil this need for appropriate teacher assessment practices to be realised.

Acknowledgement

This study was funded by the Dutch Organisation for Scientific Research (NWO-PROO 490–23–081).

Notes on Contributors

MARIEKE F. VAN DER SCHAAF is a Ph.D. student and educational researcher at the Department of Educational Sciences, Utrecht University. Her main areas of research include teacher evaluation and teaching and assessment of students' research skills. *Correspondence:* Department of Educational Sciences, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands. E-mail: m.vanderschaaf@fss.uu.nl

KAREL M. STOKKING is Professor of Education at the Department of Educational Sciences, Utrecht University. His areas of interest include the development and assessment of general and academic skills in secondary and higher education.

NICO VERLOOP is Professor of Education and Dean of ICLON Graduate School of Education, Leiden University, The Netherlands. He is immediate past president of the Dutch Educational Research Association. His major research interests are teachers' practical knowledge, learning and professional development of teachers and teacher evaluation.

REFERENCES

ARTHUR, W., WOEHR, D. J. & MALDEGEN, R. (2000) Convergent and discriminant validity of assessment center dimensions: a conceptual and empirical reexamination of the assessment center construct-related validity paradox, *Journal of Management*, 26 (4), pp. 813–835.

BERK, R. A. (1996) Standard setting: the next generation (where few psychometricians have gone before!), *Applied Measurement in Education*, 9 (3), pp. 215–235.

Cooksey, R. W. (1996) *Judgment Analysis: theory, methods, and application* (San Diego, CA, Academic Press).

Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984) A comparison of three methods for establishing minimum standards on the National Teacher examination, *Journal of Educational Measurement*, 21, pp. 113–130.

Day, C. (1999) *Developing Teachers: the challenges of lifelong learning* (London, Falmer Press).

Delbecq, A. L., van de Ven, A. H. & Gustafson, D. H. (1975) *Group Techniques for Program Planning: a guide to nominal group and delphi processes* (Glenview, IL, Scott, Foresman & Co.).

Eraut, M. E. (1994) *Developing Professional Knowledge and Competence* (London, Falmer Press).

Fitzpatrick, A. R. (1989) Social influences in standard setting: the effects of social interaction on group judgments, *Review of Educational Research*, 59 (3), pp. 315–328.

Gonczi, A. (1993) Competency based assessment in the profession in Australia, *Assessment in Education Principles Policy and Practice*, 1 (1), pp. 27–41.

Hambleton, R. K. (1997) Standard setting in criterion-referenced tests, in: J. P. Keeves (Ed.) Educational *Research, Methodology, and Measurement: an international handbook*, 2nd edn, pp. 798–802 (Oxford, Elsevier Science).

Hambleton, R. K. & Plake, B. S. (1995) Using an extended Angoff procedure for setting standards on complex performance assessments, *Applied Measurement in Education*, 8, pp. 41–55.

Huberman, M. (1995) Professional careers and professional development and some intersections, in: T. Gusky & M. Huberman (Eds) *Professional Development in Education: new perspectives and practices*, pp. 193–224 (New York, NY, Teacher College Press).

Jaeger, R. M. (1990) Setting standards on teacher certification tests, in: J. Millman & L. Darling-Hammond (Eds) *The New Handbook of Teacher Evaluation: assessing elementary and secondary school teachers*, pp. 295–321 (Newbury Park, CA, Sage).

Jaeger, R. M. (1995) Setting performance standards through two-stage judgmental Policy Capturing, *Applied Measurement in Education*, 8 (1), pp. 15–40.

Kane, M. (1994) Validating the performance standards associated with passing scores, Review of Educational Research, 64 (3), pp. 425–461.

Mehrens, W. A. (1990) Combining evaluation data from multiple sources, in: J. Millman & L. Darling-Hammond (Eds) *The New Handbook of Teacher Evaluation: assessing elementary and secondary school teachers*, pp. 322–334 (Newbury Park, CA, Sage).

Putnam, S. E., Pence, P. & Jaeger, R. M. (1995) A multi-stage dominant profile method for setting standards on complex performance assessments, *Applied Measurement in Education*, 8 (1), pp. 57–83.

Van de Pligt, J. (1997) Judgement and decision making, in: G. R. Semin & K. Fiedler (Eds) *Applied Social Psychology*, pp. 30–64 (London, Sage).

van der Schaaf, M. F., Stokking, K. M. & Verloop, N. (2003) Developing teaching content standards using a delphi method, submitted for publication.

Wheeler, P. & Haertel, G. D. (1993) *Resource Handbook on Performance Assessment and Measurement: a tool for students, practitioners, and policymakers* (Berkeley, CA, Owl Press).