DEVELOPING CONTENT STANDARDS FOR TEACHING RESEARCH SKILLS USING A DELPHI METHOD[1]

M.F. van der Schaaf, K.M. Stokking, N. Verloop

ABSTRACT The increased attention for teacher assessment and current educational reforms ask for procedures to develop adequate content standards. For the development of content standards on teaching research skills, a Delphi method based on stakeholders' judgments has been designed and tested. In three rounds, 21 stakeholders judged and revised content standards. The support for the standards increased over the rounds. The method resulted in nine content standards with a high degree of support and consensus. Qualitative analysis of stakeholders' comments and a Homals analysis showed that the stakeholders differ in the perspectives and preferences from which they judged the standards. The Delphi method proved to be an adequate procedure for developing teaching content standards based on stakeholders' judgments.

1 Introduction

The quality of education highly depends on what teachers do (the teaching tasks they perform) and on their competences to adequately fulfil these tasks. To enhance good teaching, in many countries teaching standards have been produced and discussed. Given the elusive and contested nature of what to count as good teaching, these standards can only be formulated in a relative way. It has become widely accepted that standards are shared by a community of practitioners at a certain time and place, and thus are socially situated constructs.

The need for adequate procedures for the development of standards by practitioners is urgent for at least three reasons. Firstly, in many countries, educational reforms expect teachers to change their practices towards more constructivist views on learning and teaching. Inspiring standards that can be used as a descriptive framework to rely on when dealing with the reforms can assist teachers. Since reforms can only succeed if they fit in with the teachers' beliefs,

[1] This chapter is submitted as: Van der Schaaf, M.F., Stokking, K.M., & Verloop, N. (2005). *Developing teaching content standards using a delphi method*. Manuscript submitted for publication.

everyday practices and resources, teachers at least should have a voice in developing such standards. Secondly, as constructing teaching standards is a resource for deliberation on what it means to be a good teacher (Ingvarson, 1998), this corresponds with the tendency to give teachers as a collective more responsibility in monitoring the quality and further development of their profession. However, in reality teachers still exercise little control in these matters. Thirdly, there is an increased attention for teacher assessment. Assessment can be formative, aiming at further professional growth, or summative, e.g. for certification or merit pay. Assessments are designed based on standards and assessors judge whether teachers' performances meet the standards.

Teaching standards are of two sorts: content standards and performance standards. Content standards are the criteria that conceptualise the desired tasks teachers have to fulfil in specific contexts and the competences needed to do this. Performance standards are the cut-off scores indicating to what extent teachers should meet the criteria.

In this chapter, we focus on the development of content standards. The study is part of a larger project on the assessment of social sciences teachers (economics, geography, history) in pre-university education in The Netherlands (ages 16-18), comparable with undergraduate college level in, for example, the American system. We focussed on a recent Dutch educational reform that is representative for the current change in many countries towards more constructivist views on learning that emphasize developing students' skills in an active, self-directed and collaborative way. In a number of national curricula this has led to the inclusion of research skills, both in a more restricted form, for example limited to data gathering and analysing skills (e.g. Germany, France) and in a more comprehensive way regarding the whole cycle of doing research (e.g. The Netherlands) (Steenstra, 2003).

During the last decades, different methods for generating content standards have been used (Berk, 1996), all focussed on stakeholders' discussion. Main disadvantages of such methods using face-to-face communication are interdependences between panel members, such as domination, group thinking and polarization, distorting the results and threatening their reliability (Riggs, 1983). The Delphi method is a promising alternative. This method is task-oriented, reduces social influence processes, and has a cyclic character resulting in successive refinement in the answers and generating convergence (Scheele, 1975). Whereas the Delphi method has several applications, in most cases it is used as a survey technique in which panel members respond anonymously and independently to a set of questions in successive rounds. After each round the answers are analysed and feedback is given. The aim of this procedure is to realize a structured and iterative written discussion and to foster consensus between the panel members (Linstone & Turoff, 1975).

However, the feasibility of the method for generating content standards is still largely unknown. In this study we inquire into the advantages and disadvantages of the Delphi method for the development of teaching content standards. The main questions are: How can a Delphi method be used and optimised as a procedure to develop content standards for teacher assessment in the context of educational reform? Which content standards can be set regarding teaching students research skills?

2 Validation strategies for content standards

Today, the development of content standards is commonly based on comparisons to a group or to a domain. The scores are compared with those of a well-defined group of other candidates who took the assessment before or at the same time (norm-referenced), or with a set of specified constructs derived from a properly formulated performance domain (criterion-referenced). The first approach is criticized being not clear about the relationship between the assessment outcomes and the fact that the outcomes can be interpreted in many ways. The latter approach is criticized on the premise that performance domains can be specified so precisely that items for assessing the domain could be sampled automatically and without doubt. Nevertheless, the second approach suits best the purposes of our study, because it enhances the interpretation of assessment outcomes while referencing to the domain that is assessed.

However, since assessments are used in social settings, assessment results are only relevant with a reference to a particular population. Consequently, any criterion-referenced assessment is attached to a set of norm-referenced assumptions (Angoff, 1974). This increases the ambiguous character of content standards and several researchers concluded that such standards are best based on standard setters' 'connoisseurship', that is "the ability to make fine-grained discriminations among complex and subtle qualities" (Eisner, 1991, p. 63). Nowadays this ability is also often associated with members of a community of practitioners (cf. Dylan, 1996), or, in a broader sense, to stakeholders.

Kane (1994) describes three types of strategies for demonstrating the validation of content and performance standards which are developed by stakeholders (see also chapter 4): i) validation checks based on external information, consisting of comparisons of the content standards with external information about the desired teacher practices; ii) validation checks based on internal information, concerning the consistency of the procedure used, which indicates the reliability. The word ‘indicates’ is used deliberately, because variation in panel members’ judgments does not necessarily mean that the procedure as such is not reliable, since the panel members could interpret the content standards differently; iii) procedural evidence, consisting of demonstrating the soundness of the procedure followed, including the selection of the panel members involved in the process. We included all three types of evidence in our study.

3 Method

3.1 The development of a preliminary set of content standards

We use the Delphi method to modify a preliminary set of content standards. We developed this preliminary set in a four-step procedure. The first step was the specification of the domain to be measured (Berk, 1980). Given the fact that there are several ways of good teaching and that teaching is context-bound, such a domain cannot be clearly demarcated. For that reason we used ‘domain’ only as an indicator of the information for which the standards are expected to account (Messick, 1989). We conducted a preparatory literature study to develop a general framework for describing teaching tasks. Secondly, we conducted a literature study into teaching students research skills. This was followed by an empirical study into leading edge teachers’ practices regarding developing students’ research skills (see chapter 2). Thirdly, based on the results of these studies we made an initial selection of relevant teaching tasks. Fourthly, we described a preliminary set of content standards around the tasks.

1. General framework for describing teaching tasks. Currently teaching is viewed as a complex activity in which teachers’ practices are inextricably bound up with teachers' unique cognitions, personality, and the context at hand (Shulman, 1986; Calderhead, 1996). Though teaching is highly personal, we believe that teaching as a profession at a general level can be characterized by elements that at least groups of teachers have in common (e.g. teachers who teach certain subjects to students of similar age level, as in our study) (cf. Verloop, Van Driel & Meijer, 2001).

A central element of teaching is that teaching tries to bring about learning on the part of the student (Fenstermacher & Richardson, 2000). Therefore, teachers conduct professional tasks. Reynolds (1992), after having analysed tasks from job descriptions, teachers' practical knowledge (Shulman, 1987), performance assessment systems, and duties of teachers (Scriven, 1988), synthesizes them into four domains. These include: tasks before teaching (e.g. planning instruction), tasks during teaching (instructing, coaching, assessing), tasks after teaching (e.g. reflecting on own teaching), and administrative tasks (e.g. management tasks).

Executing teaching tasks is mediated by teachers' cognitions. Putnam & Borko (1997), using Shulman (1986, 1987) and Grossman (1990), classified the content of teachers' cognitions into: general pedagogical, subject matter, and pedagogical content knowledge. Grossman (1995) added the categories of knowledge about the (school) context and the self. Especially the category of pedagogical content knowledge is viewed as a vital domain of teaching, because it refers to essential knowledge and skills that are unique to the teaching profession. It covers the translation of subject-matter knowledge into teachers' acting (Van Driel, Verloop & De Vos, 1998). In our study, pedagogical content knowledge about teaching students research skills is central: what a teacher knows, what a teacher does, and why, when teaching research skills in their subject (Baxter & Lederman, 1999).

Content standards should leave enough room to fit diverse teaching contexts, but on the other hand they should be precise enough to discriminate between suitable and unsuitable performances (Katz & Raths, 1985; Kagan, 1990). A formulation of content standards in terms of dispositions yields a middle 'size' between too extensive and too narrow descriptions. Dispositions can be seen as characteristics of teachers that summarize the trend of their intentional actions (Katz & Raths, 1985). They address the need for the likelihood that a teacher will perform in a certain way and goes beyond the idea that a teacher could have certain knowledge and skills but not employ them. Dispositions suggest what a teacher is likely to do, rather than what a he or she can do at peak performances, and therefore should be part of the content standards. Having a disposition includes the possession of relevant skills and knowledge. Dispositions also cover components of psychological nature (e.g. enthusiast, accepting, stimulating, sensitive, flexible) and moral and ethical kind (e.g. honesty, respectful, trustful) (cf. Katz et al., 1985; Reynolds, 1992; Fenstermacher et al., 2000). In this study we use the term competences to denote teacher dispositions, because both concepts are almost analogous and the former is more common.

Elaborating on this, we describe teaching tasks in terms of task domains (e.g. tasks before, during, and after teaching) and contents or cognitive domains of teaching (e.g. general pedagogical knowledge, pedagogical content knowledge, knowledge of the school context). The combination of these two dimensions generates a matrix. The cells in the matrix contain a specification of the tasks teachers have to fulfil. We formulate the standards around these tasks, in terms of teachers' dispositions to fulfil the tasks. See Figure 1.

	Knowledge domains		
Tasks	*General Pedagogical Knowledge*	*Pedagogical Content Knowledge*	*Knowledge of School Context*
Planning Developing		1. GOAL 2. ASSIGN	8b. COLLAB
Managing Instructing Coaching	3. MANAGE 6. CLIMATE	4. THINK 5. TEACH	
Assessing Reflecting	7. ASSESS 8a. REFLECT		

Figure 1. *Content standards in function of tasks and knowledge domains*

2. The domain of teaching students research skills. In a literature study Stokking & Van der Schaaf (2000) identified, based on the literature, main components of constructivist learning environments that teachers can create for their students when teaching research skills. These are goals and objectives, instruction and guidance, tasks or assignments for the students, and assessment. According to the instructional design literature, the different components of the learning environment must be consistent with one another to facilitate learning. Better results are achievable in educational settings where the curriculum and the assessment methods are aligned (Cohen, 1987). Biggs (1996) proposes the term ‘constructive alignment’ as a marriage of constructivism as a framework for instructional design with the principle of alignment. The principle of constructive alignment means that teachers are expected to be clear about the goals they pursue, to choose student-oriented instruction and guidance, to work with assignments that are suited to students’ levels of learning, to be authentic in their assessment, and to make these components fit together.

This literature study was the input for an empirical study into leading edge teachers’ practices. We executed oral interviews (n = 20) and two written surveys (n = 49; n = 165) with pre-university teachers in social science subjects (economics, history, geography) and natural sciences (physics, biology). Also teacher submitted research assignments and associated information (teachers’ goals, ways of coaching, assessment procedures and criteria) (n = 54). The study showed that the main components as revealed by the literature are indeed important aspects of teachers’ practices. Furthermore, the study revealed additional aspects that play important roles in teaching students research skills: the collaboration with colleagues (e.g. making arrangements with regard to the teachers’ goals, students’ assignments and the assessment) and the bottlenecks teachers experience when handling research assignments (e.g. due to lack of time for coaching and assessment, lack of modern media at school, and lack of information resources and other material needed, the support some students get from others such as their parents, and fraud) (see Stokking, Van der Schaaf, Jaspers & Erkens, 2004).

3. Initial selection of teaching tasks. We selected teaching tasks that emerged to be the most relevant in the domain of teaching students research skills, and that fitted in the matrix of tasks and knowledge domains. Also, the tasks as an indication guarantee for being consistent constructs had to be measured reliably in the surveys. The selected tasks include: setting goals (Cronbach’s alpha .62); selecting assignments (.88); instruction and coaching (.86); using different coaching approaches (.63); assessing students’ research skills (.83); collaborating with colleagues (.93). In addition, we selected tasks regarding management of the research assignments. Finally, we added the establishment of a safe learning climate. In sum, we derived a provisional list of these tasks: 1. Formulating goals; 2. Developing assignments; 3. Managing students working on their assignments; 4. Thinking of teaching (preparation); 5. Teaching research skills (instruction); 6. Creating a positive pedagogical climate; 7. Assessing students' research skills; 8. Reflecting (8a) and collaborating (8b) with colleagues.

4. Initial description of content standards. Relying on our data, we closely examined each category and made an initial description per task. For example, our data showed that teachers pose requirements to the research assignments they select. The teachers in general want the assignments to fit close to the students' level, to contain enough subject content, to be realizable in a certain amount of time, to be sufficiently challenging, and to produce testable results, and also offer the students options, and to be approachable in various ways. Furthermore, about 40% of the teachers almost always let the students work on a research assignment in pairs or small groups. We summarized this information in our initial description of a content standard concerning the development of assignments.

Each content standard was thoroughly described, in a half to full page per standard. We described the content standards in a document of 7 pages, consisting of a description in terms of dispositions and a more detailed description to direct the interpretation of the standards (Ryan & Kuhs, 1993). The content standards that were still rather broad were divided into two or more indicators (see Table 2). The content standards were converted into a questionnaire that formed the input to our Delphi study.

3.2 The sampling of panel members

In order to produce high quality standards several conditions have to be fulfilled. (a) The selected stakeholders should have (practical) knowledge of the teaching research skills in social science subjects in pre-university education and about everyday teaching practices. (b) The stakeholders have to agree with the way in which the content standards to be set will be used in the assessment of teachers (Putnam, Pence & Jaeger, 1995). (c) Although reproducible content standards may be accomplished with 5 to 10 judges (Norcini, Lipner, Langdon & Strecker, 1987), a larger group will increase the chance that a variety of expertise, experiences, and perspectives is represented. The Delphi method is suitable for 10 to 50 participants (Turoff, 1975).

Heterogeneous groups with different backgrounds and various perspectives typically develop highly acceptable and valuable results (Norcini & Shea, 1997), but if a group is not constructed with sufficient care, the quality of the representation of the population in the group and the resulting quality and acceptability of the content standards will be reduced (Linstone & Turoff, 1975; Sackman, 1975; Clayton, 1997).

Preparing the selection process of the participants, we involved groups with a considerable stake in the results of the study (Clayton, 1997): teachers in geography, history, and economics (n = 3), educational researchers (n = 3), government employees (n = 4), lobbyists (n = 2), pedagogical content experts in geography, history, and economics (n = 4), teacher educators (n = 4), and school principals (n = 2). These individuals were interviewed about the assessment of teacher competences in teaching students research skills. When asked which persons or organizations should participate in the process of developing content standards for summative teacher assessment, most interviewees mentioned practising teachers, followed by external experts such as teacher educators and pedagogical content experts.

We used the following procedure to make up the Delphi-panel. From a random sample of 115 schools we approached the heads of the geography, history, and economics departments and their school principals with information about the study (goals and planning), a description of the judgment task, a request for participation, and a detailed intake form. The intake form contained questions on general characteristics (sex, age, position, years of experience), opinions on students' research in social science disciplines, and their experience with teacher assessment (Putnam et al., 1998).

Fifteen respondents were willing to participate: 3 school heads and 12 experienced teachers: 3 in geography, 4 in history, 3 in economics, and 3 in social science subjects who were also teacher trainers. This low response can be explained by the demanding character of the study in terms of motivation, time, and expertise. In addition, 2 educational researchers, 3 pedagogical content specialists, and a lobbyist were asked to participate. Thus, 21 individuals participated in the Delphi study.

3.3 Data collection

The stakeholders received instructions, including a description of the aims of the study, of the way the content standards to be set will be used, and of teaching students research skills.

The number of rounds that is needed depends on how stable the judgments of the panel members are and on how quickly they reach an acceptable level of (statistical) consensus (Erffmeyer, Erffmeyer & Lane, 1996; Smith & Simpson, 1995). Often, three rounds are enough. Therefore, in three monthly cycles the panel members judged the content standards and indicators. Their judgments were based on the following four aspects: content relevance, thoroughness of formulation, clarity of formulation, and correspondence of the content standards with everyday teaching practice. Firstly, they answered open questions about the proposal as a whole. Secondly, for each content standard they answered half-open questions on the four aspects. Thirdly, for each content standard and each indicator they were asked to rate the four aspects on a 4-point Likert scale. Fourthly, in order to gain insight into the way the panel members interpret the content standards, they were asked to comment on the ratings they made. Fifthly, they were asked to rank the content standards in order of importance.

To improve the content standards, in each round the panel members were asked for suggestions. After each round the standards were revised, based on their' judgments and comments, after analyses by two researchers. The panel members received written feedback concerning the means and standard deviations of the ratings of the whole group, a summary of the comments made by the other panel members, and an overview of the conclusions per round. In the third round the panel members were also asked to evaluate the Delphi method itself. They answered open and closed questions about their experiences and their level of satisfaction with the method.

3.4 Data analysis

After each round a key issue was the decision whether particular content standards should be accepted, revised, or deleted. This decision was based on the validation evidence, and directed by the panel members' ratings of the content standards and indicators on the four aspects: content relevance, thoroughness of formulation, clarity of formulation, and correspondence with everyday teaching practice. For the analysis of the panel members’ judgments per round and across the rounds we formulated the following guidelines.

Firstly, we wanted to know whether the group supports, opposes, or is ambivalent towards the content standards. We used the mean to indicate whether the panel members support the content standards (4 = strongly supports, 3 = weakly supports, 2 = weakly opposes, 1 = strongly opposes the content standard).

Secondly, we wanted to analyse the degree of consensus between the judgments. We divide the concept consensus into statistical consensus (percentage agreement) about the content standards and substantial consensus about underlying perspectives. We allow panel members to have different valid perspectives which all can contribute to the definition on the content standards. Following the results of a previous Delphi study by De Loe (1995), beforehand we distinguished four levels of agreement: high (70% of the ratings in one category or 80% in two contiguous categories); medium (60% of ratings in one category or 70% in two contiguous categories); low (50% of ratings in one category or 60% in two contiguous categories); none (less than 60% in two contiguous categories). However, consensus can only be assumed if judgments tend to be in one direction only (Sackman, 1975). The skewnesses of the distributions of the ratings were computed to check on symmetry and to check whether the panel members’ judgments tended to be in one direction.

Thirdly, the interrater reliability was checked by computing jury alphas, and the consistency between items was examined by computing Cronbach’s alphas.

Fourthly, our purpose was to revise the content standards based on the judgments and comments of the panel members until a list emerged that suited the majority of panel members. To test whether the succeeding rounds resulted in increasing support, we compared the ratings between rounds by performing paired sample T-tests. To check whether the range in the answers decreased, we performed paired sample T-tests on the deviation scores.

Fifthly, to gain insight into the most important content standards, the panel members were asked to rank these standards in order of importance. The consistency in the rankings between the rounds was estimated by calculating correlations.

Finally, we analysed stakeholders’ written comments for emerging themes and categories. The data of the third round was analysed by homogeneity analysis (Homals) to see whether the panel members could be grouped according to their rankings. Homals projects the differences between panel members concerning their preferences into distances between points, representing the panel members, in a two-dimensional plot.

4 Results

4.1 Results per Round

First Round

Results – overall judgment. The panel was asked whether the content standards satisfactorily covered the subject of our study, teaching students research skills. More than half of the panel members thought this to be the case (57%). Only 10% did not agree. Approximately one-third did not directly answer the question, but made some specific remarks. Most noticed that some content standards overlapped. Next, the panel members answered questions about the formulation of the content standards. More than half of them thought the content standards were clear (57%). Several panel members noted that the list as a whole was quite detailed, but judged the formulation of separate content standards as too global or too abstract. Finally, the panel members were asked whether they thought it possible to check whether teachers met the content standards. Almost all agreed (90%).

Results per content standard. Table 1 gives an overview of the mean ratings per content standard. The panel supported the relevance of most content standards. It weakly supported the thoroughness and clarity of the formulation, and did not support the correspondence with everyday teaching practice of most content standards. The degree of consensus varied strongly between the content standards, but for most standards it was mediocre.

Revision. The revision mainly consisted of reformulating the content standards in a less abstract manner. This was partly realized by adding illustrations and by categorizing the standards more clearly. In addition, the standard ‘Reflecting on the program and on personal actions, and collaborating with colleagues in teaching students research skills’ was divided into two standards. Note that in rounds 2 and 3 the content standard 'Reflecting on the program and on personal actions, and collaborating with colleagues' was divided into two content standards. The positive answers as to the relevance of the standards indicated that it was not necessary to repeat the relevance question, and so we omitted this question in the next rounds.

Second Round

Results – overall judgment. The panel was asked open-ended questions about whether they thought the reformulation of the content standards to be an improvement. Almost three-quarters of the panel members (71%) thought this to be the case. In summary, the panel thought that the content standards resulted in a better survey and were formulated more realistically. Most standards were judged as being to the point; some standards were seen as too elaborate.

Results per content standard. In the second version, 'Collaborating with colleagues' had been added as a new, separate content standard. Panel members were asked whether it was sensible to distinguish this standard separately. The majority thought this to be so to a reasonable (34%) to strong (52%) degree. Table 1 shows that the panel members judged the content standards in the second round to be, on average, formulated thoroughly and clearly to a reasonable to strong degree. The correspondence with everyday practice was judged as reasonable.

Revision. The most important conclusion from the second round was that the content standards had to be formulated in an even more condensed and concrete form. The resulting changes concerned more compact formulations of the content standards and the addition of a 'shortlist' with a summary of the content standards.

Third Round

Results – overall judgment. The panel was asked whether they thought the reformulations were an improvement. Three-quarters (76%) agreed and 14% partly agreed. The formulations were evaluated as more concise, clearer, a better match to the practice of teaching, more logically constructed, less redundant, and linguistically more correct. Two-thirds of the panel (62%) thought all elements were to the point.

Results per content standard. Table 1 shows that the content standards on average were judged to be thoroughly and clearly formulated and corresponding to everyday practice to a reasonable or strong degree. The average scores for thoroughness and clarity were all above 3,3. The average scores for corresponding to practice ranged from 3,0 to 3,4. On all aspects with regard to all standards, the degree of consensus was high.

Table 1. *Support and consensus of the content standards in rounds 1-3 (n = 21)*

	First round				Second round				Third round			
	Support		Consensus		Support		Consensus		Support		Consensus	
	m	sd	a	sk	m	sd	a	sk	m	sd	a	sk
1. GOAL												
Relevance	3,7	.6	H	-1.8								
Formulation (t)	3,3	.9	H	-1.0	3,7	.5	H	-0.9	3,8	.4	H	-1.3
Formulation (c)	2,9	.9	M	-0.5	3,4	.7	H	-0.7	3,7	.5	H	-0.9
Correspondence	2,5	.9	L	0.6	3,0	.9	M	-0.5	3,2	.8	H	-1.2
2. ASSIGN												
Relevance	3,7	.6	H	-1.7								
Formulation (t)	3,3	.8	M	-0.4	3,7	.5	H	-0.8	3,6	.6	H	-1.3
Formulation (c)	3,2	.6	H	-0.9	3,3	.9	H	-1.1	3,4	.6	H	-0.3
Correspondence	2,6	.9	N	0.3	2,9	.9	M	-0.7	3,3	.7	H	-0.8
3. MANAGE												
Relevance	3,5	.5	M	0.1								
Formulation (t)	3,3	.6	M	0.0	3,5	.6	H	-0.6	3,4	.6	H	-0.5
Formulation (c)	3,3	.6	M	-0.2	3,5	.5	H	0.2	3,6	.5	H	-0.3
Correspondence	3,0	.8	N	0.3	3,0	.8	M	-0.6	3,2	.8	H	-0.6
4.THINK												
Relevance	3,1	.7	M	-0.6								
Formulation (t)	3,0	.6	H	0.0	3,8	.4	H	-1.3	3,7	.6	H	-1.6
Formulation (c)	2,7	.7	N	0.7	3,6	.5	H	-0.4	3,5	.8	H	-1.9
Correspondence	2,1	.7	N	0.0	3,1	.9	M	-0.8	3,2	.8	H	-0.7
5. TEACH												
Relevance	3,4	.6	H	-0.4								
Formulation (t)	3,1	.7	M	-0.1	3,5	.5	H	0.0	3,6	.5	H	-0.3
Formulation (c)	3,1	.7	M	-0.1	3,5	.5	H	0.0	3,6	.5	H	-0.3
Correspondence	3,0	.9	L	-0.1	3,0	.8	H	-1.4	3,2	.7	H	-0.6
6. CLIMATE												
Relevance	3,6	.6	M	-1.3								
Formulation (t)	3,1	.8	L	-0.5	3,5	.7	H	-1.1	3,5	.7	H	-1.0
Formulation (c)	3,3	.7	M	-1.0	3,5	.6	H	-0.9	3,6	.5	H	-0.6
Correspondence	2,9	.6	M	-0.7	3,0	.5	M	0.1	3,4	.6	H	-0.5
7. ASSESS												
Relevance	3,3	.8	M	-0.6								
Formulation (t)	3,1	.7	M	-0.1	3,4	.6	H	-0.8	3,4	.5	H	0.3
Formulation (c)	3,3	.8	M	-0.4	3,5	.7	H	-0.6	3,5	.7	H	-1.9
Correspondence	2,4	.8	H	-0.3	2,8	.8	H	-0.2	3,1	.8	H	-0.7
8. REFLECT												
Relevance	3,3	.7	H	-0.4								
Formulation (t)	3,1	.7	M	-0.2	3,6	.5	H	-0.4	3,6	.6	H	-1.3
Formulation (c)	3,1	.7	M	-0.1	3,6	.6	H	-1.0	3,7	.7	H	-2.8
Correspondence	2,1	.6	H	-0.1	2,8	1.0	L	-0.4	3,2	.8	H	-1.1
9. COLLAB												
Formulation (t)					3,3	.7	H	-0.4	3,7	.5	H	-1.0
Formulation (c)					3,3	.6	H	-0.2	3,7	.6	H	-1.8
Correspondence					2,7	1.0	L	-0.5	3,0	.9	H	-0.6

Note. m = mean (1 = weak support, 4 = strong support); a = agreement (H = high, 70% of the ratings in one category or 80% in two contiguous categories; M = medium, 60% of ratings in one category or 70% in two contiguous categories; L = low, 50% of ratings in one category or 60% in two contiguous categories; N = none, less than 60% in two contiguous categories). sk = skewness; t = thorough; c = clear.

4.2 Interrater consistency and scale reliability

The panel as a jury was rather consistent. Table 2 shows the Jury alphas in the third round per content standard and per indicator. Overall, the agreement proved to be sufficient. The panel members only disagreed on standard 6 (CLIMATE). The Jury alphas in the three rounds for all content standards and indicators together were .53 (first round), .72 (second round) and .65 (third round).

Table 2. *Internal consistencies (Cronbach's alpha) and interrater consistencies (jury alpha) in panellists' judgments of the content standards and indicators, third round (n = 21)*

		Cronbach's alpha	Jury alpha
1.	GOAL	.67	.87
2.	ASSIGN	.56	.69
a)	selecting short-term goals	.78	.87
b)	selecting appropriate contents	.82	.91
c)	selecting an appropriate form	.58	.78
3.	MANAGE	.73	.62
a)	structuring time and room, making resources accessible	.72	.45
b)	offering insights into working and assessment procedures	.73	.74
4.	THINK	.75	.71
5.	TEACH	.74	.82
6.	CLIMATE	.63	.06
7.	ASSESS	.06	.85
a)	explicitly stating purposes of the assessment	.90	.47
b)	choosing an adequate assessment model	.83	.44
c)	conducting fair assessment practices	.71	.74
d)	communicating inferences and results to students	.62	.67
8.	REFLECT	.84	.84
9.	COLLAB	.61	.89

The judgments about the thoroughness and clarity of formulation and the correspondence with teaching practice became more consistent across the rounds; see Table 3. The jury alphas were satisfactory in each round and increased over time. The relevance was only judged in the first round and the alpha was moderate, which is obvious for this aspect in a first round.

With regard to each content standard and indicator the panel members gave judgments on the relevance, thoroughness of formulation, clarity of formulation, and correspondence with everyday teaching practice. In each round regarding each standard and indicator, the judgments on these aspects were analysed for consistency and to ascertain whether, in the case of several indicators, the indicators made up a sufficiently reliable scale or had to be distinguished separately. We limit ourselves to reporting on the third round. Table 3 shows the results of these item analyses.

Table 3. *Internal consistency (Cronbach's alpha) of the aspects across content standards per round (n = 21)*

	First round			Second round			Third round		
	m	sd	α	m	sd	α	m	sd	α
Relevance	3,33	.45	.61						
Formulation (thorough)	3,21	.41	.82	3,55	.33	.88	3,63	.29	.91
Formulation (clear)	3,10	.43	.68	3,43	.35	.91	3,61	.38	.86
Correspondence	2,46	.82	.82	2,93	.76	.90	3,24	.60	.94

Cronbach’s alphas of all standards without separate indicators (standards 1, 4, 5, 6, 8, 9) were sufficiently high. Amongst the standards made up of different indicators, standard 3 (MANAGE) formed a sufficiently reliable scale. Standard 2 (ASSIGN) was only a weak scale. It seems sensible to distinguish its indicators separately but we nevertheless decided to keep the scale. The alpha of standard 7 (ASSESS) was very low, while the separate indicators had sufficiently high alphas.

4.3 Results across the rounds

The panel’s support increased across the rounds. The panel members tended to rate the standards in the successive rounds as being more thoroughly and clearly formulated and more corresponding with practice (see Table 1). Paired sample T-tests between the ratings of the first and third rounds showed that the standards 1 (GOAL), 2 (ASSIGN), 3 (MANAGE), 4 (THINK), 5 (TEACH), and 6 (CLIMATE) were judged significantly higher on ‘Thorough formulation’ ($p < .05$). The standards 1, 4, 5, and 6 were rated significantly higher on ‘Clear formulation’ ($p < .05$). The standards 1, 2, 3, 4, 6, and 7 (ASSESS) were rated significantly higher on ‘Correspondence with everyday teaching practice’ ($p < .05$). The T-tests on the deviation scores were not significant, so the range in the judgments did not change.

4.4 The panels’ underlying asssumptions and perspectives

The written comments of the panel members in the three rounds were transcribed verbatim, resulting in about 40 pages of text. These were then qualitatively analysed. We now summarize the most important perspectives of the panel members in judging the content standards. The abbreviation P (1-21) stands for ‘panel member’, R (1-3) for ‘round’, S (1-4) for ‘support score’, and (-) for ‘missing’; thus, the reference P3, R2, S4 stands for panellist 3 in round 2 giving support 4 (strong support).

1. Selecting yearlong goals for students (GOAL). The panel members thought this content standard to be relevant, for different reasons: because it is important for students to have a cumulative curriculum, and because collaboration between colleagues is important, giving the teacher focus and legitimacy in instruction, coaching, and assessing. The correspondence with everyday practice was judged less high, because teachers operate pragmatically, owing to lack of time and dependence on collaboration within the school, e.g. "Formulated very idealistically. Should be attainable, but it depends also on the department and the school agreements" (P18, R3, S-).

2. Choosing an appropriate assignment (ASSIGN). In the first round almost all panel members thought this standard to be relevant, because "Everything depends on having an adequate assignment" (P11, R1, S4; P13, R1, S4; P18, R1, S4).

Otherwise, the practices differ, e.g. "The student's contribution is made to heavy, in practice the teacher will decide how the assignment will be shaped" (P2, R3, S3).

3. Preparing and managing students working on their assignments (MANAGE). Panel members who judged this content standard in the first round as highly relevant, thought organization to be a condition for working on the assignment, e.g. "Classroom management is very important in coaching independent research" (P20, R1, S4). Panel members who provided the judgment 'rather relevant', gave two considerations. The content standard was thought to be not only the teacher's task but also the student's, e.g. "Of course these things are relevant, but part of it can be managed by the students themselves" (P6, R1, S3), or other content standards were judged as more important, e.g. "A little less relevant than other indicators, because these matters are only organizational and I guess have less influence on learning" (P19, R1, S3). The panel members who thought that students can manage things themselves judged the correspondence of the content standard to everyday practice to be less high in the first and second round, because the standard assumed to too great an extent that the teacher was in control; in the third round, when the standard was reformulated to give the students more control, they thought the standard was more recognizable, e.g. "This is roughly how things are going on in our school" (P6, R3, S3). However, some panel members judged the formulation in the third round to be less recognizable, e.g. "Much too optimistic and formulated far from reality" (P2, R3, S-).

4. Previously thinking of and selecting teaching strategies that support the development of research skills of students (THINK). Several panel members thought this content standard to be relevant because it is a condition for teaching students, e.g. "Otherwise many students will fail" (P15, R1, S4). Some panel members judged the content standard to be less relevant because it is so obvious or because it goes together with other content standards, e.g. "This standard belongs to teaching students research skills" (P18, R1, S2). In the first round the correspondence with everyday practice was judged as moderate, because it assumes quite a lot of coordination in the school or because teachers are not so far, e.g. "I am convinced that most colleagues are not able to do this adequately. For sure, it should be done this way" (P2, R3, S2).

5. Teaching students research skills (TEACH). Most panel members thought this content standard to be relevant because it is the kernel of teaching: "If this is not what it is all about, what is it?"(P12, R1, S4). Because of lack of time the correspondence with everyday practice was judged to be in general somewhat less high.

6. Creating a positive pedagogical climate (CLIMATE). The panel members who judged this content standard as highly relevant thought that creating an adequate climate is an outstanding condition for good teaching and guidance, e.g. "It is a general condition for education" (P8, R1, S4; P9, R1, S4; P21, R1, S4). Some panel members remarked that this standard is a general one, and not specific to the promotion of reform 'independent research'. For this reason a few panel members judged the content standard as somewhat less relevant. Two panel members gave the warning that the pedagogical climate can become an umbrella construct, e.g. "This could become endless" (P22, R2, S-). Finally, several panel members indicated that the climate in class depends on the school climate and the teacher's personality: "This standard depends very much on the school climate and the teacher personality; teachers will differ in this to a great extent" (P2, R3, S3); "It is important to respect differences between teachers. Not everybody will be able to create an adequate climate in the same way." (P13, R1, S3).

7. Adequately assessing research skills of students (ASSESS). Several panellists thought "Assessing is an important component in the learning process of students" (P4, R1, S3; P6, R1, S4). When judging the correspondence with everyday practice, many added that teachers have problems with assessing students.

8. Reflecting on the program and on personal actions in teaching students research skills (REFLECT). Most panel members thought this content standard to be relevant but pointed out that teachers need training because they do not reflect as a matter of course. In addition, the school climate is important here. Consequently, the correspondence with everyday practice is judged to be rather low. "Lack of time and facilities, it is a school's task" (P1, R2, S2). "Does not happen explicitly, demands consultation and training" (P15, R2, S2). In contrast to this, a couple of panellists thought this content standard so obvious that they doubt whether it should be made explicit.

9. Collaborating with colleagues (COLLAB). Although most panel members judged this content standard to be relevant, e.g. "This has to be accentuated, the teacher is a team player" (P1, R2, S3), a few thought the standard to be less relevant because it is obvious and overlaps with other standards: "Collaboration with colleagues is obvious. This is not an independent standard!" (P10, R2, S1). In addition, opinions differed as to the correspondence with practice: "This is in our school as yet very moderately developed, people are stuck in their own discipline" (P6, R2, S2) versus "Recognizable, but it stays dependent on the employment conditions in education and on the colleagues one has to collaborate with" (P21, R3, S4).

The panel members' comments show that their judgments are based on different experiences in their own (educational) practice. Important themes that arise are the feasibility and advisability of collaboration between teachers, the extent to which students can work independently, the fact that content standards are interwoven, and the need to train teachers with regard to some standards. Although the panel members in general judge all content standards to be important, they prefer some content standards above others.

In order to map the panel members' preferences we asked them in each round to rank the content standards. Based on the rankings in the third round we constructed a data matrix with 20 cases (20 responding panel members) and 36 variables (all pairs of 9 content standards [9(9-1)/2]). The rankings were converted into pairwise comparisons, scoring 1 if the first standard was preferred to the second, and 2 if otherwise. In searching for clusters in the answers of the panel members, a homogeneity analysis (Homals) was carried out. The result was a mapping of the panel members on two dimensions with eigenvalues of .302 and .178 (total fit .48).

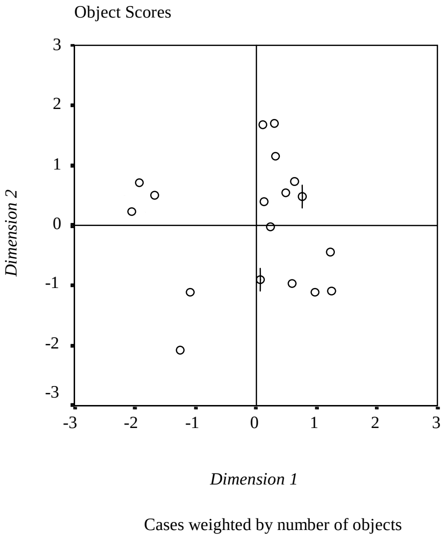

Figure 2. *Panellists' preferences for the content standards (Homals) (n = 20)*

The eigenvalues are moderate, which means that the panel members cannot be easily distinguished. They fall into four clusters, each with its own preferences. The clusters turned out not to be related to the background and positions of the panel members. Although it is sensible not to over interpret the results, Figure 2 shows that the panel members with low scores on the first dimension (n = 5) prefer the content standards 1 (GOAL) and 9 (COLLAB) above all other standards. This dimension identifies panel members who perceive teachers as team players. Panel members with low scores on the second dimension (n = 8) prefer content standards related to preparing for students' independent work, such as 2 (ASSIGN) and 3 (MANAGE), over content standards that concern the teaching process itself, such as 5 (TEACH) and 6 (CLIMATE).

4.5 The panel's evaluation of the Delphi method

At the end of the study the panel members evaluated the Delphi method (questions on a 4-point scale from 1 ‘not or to a low degree’ to 4 ‘to a high degree’, Cronbach’s alpha .73). The panel members did not need more Delphi rounds to formulate the content standards even more sharply (mean 1.2; standard deviation .70), and thought additional rounds would not add much to the results (3.6; .80). They thought the Delphi procedure yielded enough (2.7; .73), and were not much in agreement with the proposition that better methods than Delphi exist to develop content standards (2.2; 1.05). The panellists judged the questions in the questionnaires in the successive rounds to be clear (3.7; .56), and thought these to be adequate to communicate their comments on the content standards (3.1; .57). In giving their judgments in the second and third rounds, the panel members used to a reasonable degree the feedback they received from the previous rounds (2.5; .68). Finally, the Delphi method was evaluated to a reasonable degree to be a pleasant way to judge and revise content standards (2,6; .86), but at the same time judging the content standards had been experienced to be rather difficult (2.8; .87). The overall evaluation of the Delphi method (the scale mean) did not correlate significantly with the judgments of the content standards in the third round. This suggests that differences between the panel members in their evaluation of the method did not result in differences in their judgments of the proposed content standards.

5. Conclusions

An increased attention for teacher assessment and large-scale educational reforms ask for procedures to develop adequate content standards. The Delphi method is a promising alternative, because it has a task-orientated focus, reduces social influence processes, and has a cyclic character that results in successive refinement and generates convergence (Scheele, 1975). However, the feasibility of the method of generating content standards is still largely unknown. We tested the method in a small-scale standard setting study and summarize the most important merits and restrictions.

Firstly, the method makes it possible to gradually improve content standards by an iterative process. The method resulted in nine content standards that are considered by the panel members to be highly relevant, thoroughly and clearly formulated and corresponding well with everyday teaching practice. The results show satisfactory interrater and interitem consistencies.

Secondly, although the ratings became more positive across the rounds, it is not fully clear what caused this change. It may be attributed to the iterative development of the content standards using the critical remarks of the panel members, to the feedback given, to their familiarization with the content standards (a learning effect), or even to the fact the panel members have simply confirmed the majority judgment (Greatorex & Dexter, 2000). This phenomenon needs further exploration, for example by longitudinal studies into possible changes in the content of panellists' cognitive representations of the content standards during the standard setting process.

Though in general statistical consensus was reached, it was difficult to reach a high level of consensus on the aspect 'Correspondence with everyday teaching practice'. The qualitative analysis of the additional remarks indicates that this could have resulted from the very different (school) practices of each panellist. Another possible explanation lies in the broad formulation of the content standards. In the questionnaires the content standards were illustrated with a few examples, but intentionally they were unrelated to a specific classroom setting. For that reason the panellists may vary in the teaching practice they connect the content standards to, and this may result in the relative less consensus. An alternative would be to illustrate the content standards using critical incidents or using more examples of practical classroom situations.

Fourthly, two important themes on which the panel members differed in their interpretations of the content standards were the feasibility and desirability of collaboration between teachers, and the degree of independence by which students can be expected to work on research assignments. It is no surprise that analysing the preference data with Homals, the panel members are divided into four clusters that differ in just these two dimensions. So, the method shows the existence of differences between panel members' preferences. However, the method does not clearly explain what causes these differences. For that reason the Delphi method needs to be supplemented. Interviews or other face-to-face methods in which the panel members' explain their underlying perspectives may be helpful as additional methods.

As expected, our study shows that the content standards can be interpreted in different ways. Judgments and preferences are likely to be explained by panel members' backgrounds such as specialities and professional skills (Plake, Melican & Mills, 1991). In our study however, the panel members' judgments and preferences are not related to their professional position nor to their discipline. A remaining explanation is that our panel members work in different (educational related) organizations. Since it is often suggested that individuals from similar backgrounds are likely to reach agreement most readily (Pula & Huot, 1993), the Delphi method may be more suitable for internal standard setting processes.

Sixthly, although we selected a broad group of stakeholders, the resulting content standards are not based on an exhausting overview of possible perceptions. So, as expected, the meanings of content standards strictly spoken can only be defined relative to the underlying perceptions of the stakeholders that participated in our study. Since the precise meaning of the content standards becomes important when using the standards (for example in assessment situations), the persons concerned (e.g. assessors and assessees) at least should under scribe the meaning of the content standards as formulated.

Finally, the panel members judged a preliminary set of content standards already having a certain structure; they were not invited to propose a totally different set of content standards. Although the results cannot be separated from the input for the Delphi method (the questionnaire used), we tried to reduce these influences by a careful selection of content standards based on preparatory literature study and empirical research into teachers' practices, and by adjusting the questionnaire after each round only after analyses by two researchers. A pitfall is the enormous investment of time needed.

The next question is to what degree teachers have to meet the content standards for a satisfactory result. In order to be useful for teacher assessment, the content standards need to be accompanied by performance standards. Therefore, the content standards developed in this study have been used as input for a study into the development of performance standards (Van der Schaaf, Stokking & Verloop, 2003).

The need for suitable methods for developing content standards will remain for some time. In this study we wanted to make a contribution by studying the possibilities of the Delphi method as a content standard setting method. More small-scale evaluations are needed, to further improve the quality and feasibility of standard setting methods.

REFERENCES

Angoff, W.H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, 92 (summer), 2-5, 21.

Baxter, J.A., & Lederman, N.G. (1999). Assessment and measurement of pedagogical content knowledge. In J. Gess-Newsome & N.G. Lederman (Eds.), *Examining pedagogical content knowledge. The construct and its implications for Science education*, (pp. 147-161). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Beijaard, D., & Verloop, N. (1996). Assessing teachers' practical knowledge. *Studies in Educational Evaluation*, *22* (3), 275-286.

Berk, R.A. (Ed.) (1980). *Criterion-referenced measurement. The state of the art*. Baltimore/London: The Johns Hopkins University Press.

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

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher education*, 32, 347-364.

Calderhead, J. (1996). Teachers: Beliefs and knowledge. In D. C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology.* 709-725. New York: Macmillan.

Clayton, M.J. (1997). Delphi: a technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology*, *17* (4), 373-386.

Cohen, S.A. (1987). Instructional alignment: searching for a magic bullet. *Educational Researcher*, 16 (8), 16-20.

De Loe, R.C. (1995). Exploring complex policy questions using the policy Delphi. A multi-round, interactive survey method. *Applied Geography*, *15* (1), 53-68.

Dylan, W. (1996). *Construct-referenced assessment of authentic tasks: alternatives to norms and criteria*. Paper presented at the 24th Annual Conference of the International Association of Educational Assessment Testing and Evaluation: Confronting the Challenges of Rapid Social Change, Barbados, may 1998.

Eisner, E.W. (1991). *The enlightened eye. Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan Publishing Company.

Erffmeyer, R.C., Erffmeyer, E.S., & Lane, I.M. (1986). The Delphi Technique: an empirical evaluation of the optimal number of rounds. *Group & Organization Studies*, *11* (1-2), 120-128.

Fenstermacher, G.D. & Richardson, V. (2000). *On making determinations of quality in teaching*. A paper prepared at the request of the Board on International Comparative Studies in Education of the National Academy of Sciences, University of Michigan, Ann Arbor.

Greatorex, J. & Dexter, T. (2000). *An accessible analytic approach for investigating what happens between the rounds of a delphi study*. Journal of Advanced Nursing, 32 (4), 1016-1024.

Grossman, P.L. (1990). *The making of a teacher. Teacher knowledge and teacher education*. New York: Teachers College Press.

Grossman, P.L. (1995). Teachers' knowledge. In Anderson, L.W. (Ed.), *International Encyclopaedia of teaching and teacher education*. Second edition. Oxford: Pergamon.

Ingvarson, L. (1998). Teaching standards: foundations for professional development reform. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *International handbook of educational change. Part two* (pp. 1006-1031). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kagan, D.M. (1990). *Ways of evaluating teacher cognition: inferences concerning the Goldilocks Principle.* Review of Educational Research, 60 (3), 419-469.

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

Katz, L.G., & Raths, J.D. (1985). *Dispositions as goals for teacher education*. Teaching & Teacher Education, 1 (4), 301-307.

Linstone, H.A. & Turoff, M. (Eds.) (1975). *The Delphi method: techniques and applications,* (pp. 37-71). London: Addison-Wesley Publishing Company.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement,* (pp. 13-103). New York: Macmillan.

Norcini, J.J., Lipner, R.S., Langdon, L.O., & Strecker, C.A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24* (1), 56-64.

Norcini, J.J., & Shea, J.A. (1997). The credibility and comparability of standards. *Review of Research in Education, 17*, 3-29.

Plake, B.S., Melican, G.J., & Mills, C.N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: issues and practice, 10* (2), 15-16, 22, 25.

Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), Validating holistic scoring for writing assessment. *Theoretical and empirical foundations* (pp. 237–265). Cresshill, NJ: Hampton Press.

Putnam, R.T., & Borko, H. (1997). Teacher learning: implications of new views of cognition. In B.J. Biddle, T.L. Good, & I.F. Goodson (Eds.), *International handbook of teachers and teaching, 2,* (pp. 1223-1296). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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)*,* 57-83.

Reynolds, A. (1992). Getting to the core of the apple: a theoretical view of the knowledge base of teaching. *Journal of Personnel Evaluation in Education, 6,* 41-55.

Riggs, W.E. (1983). The Delphi Technique. An experimental evaluation. *Technological forecasting and social change, 23,* 89-94.

Ryan, J.M. & Kuhs, T.M. (1993). Assessment of preservice teachers and the use of portfolios. *Theory into Practice, 32,* 75-81.

Sackman, H. (1975). Delphi Critique. Expert opinion, forecasting, and group processes. Massachusetts: Lexington Books.

Scheele, D.S. (1975). Reality construction as a product of Delphi interaction. In H.A. Linstone & M. Turoff (Eds.), *The Delphi method: techniques and applications*, (pp. 37-71). London: Addison-Wesley Publishing Company.

Scriven, M. (1988). Evaluating teachers as professionals: The duties-based approach. In Stanley, J. J., & Popham, W. J. (Eds). *Teacher evaluation: Sixprescriptions for success* (pp. 110-142). Alexandria, VA: ASCD.

Smith, K.S., & Simpson, R.D. (1995). Validating teaching competencies for faculty members in higher education: a national study using the Delphi method. *Innovative Higher Education, 19* (3), 223-234.

Shulman, L.S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher, Feb. 1986,* 4-14.

Shulman, L.S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Review, 57* (1), 1-22.

Steenstra, C. (2003). *Geography's contribution to general education. An international comparative study*. (Doctoral dissertation). Delft: Eburon.

Stokking, K., Van der Schaaf, M., Erkens, G., & Jaspers, J. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal, 30* (1), 93-116.

Turoff, M. (1975). The policy Delphi. In H.A. Linstone, & M. Turoff (Eds.), *The Delphi method: techniques and applications,* (pp. 84-100). London: Addison-Wesley Publishing Company.

Van der Schaaf, M.F., Stokking, K.M. & Verloop, N. (2003). Developing performance standards for teacher assessment by policy capturing. *Assessment and Evaluation in Higher Education, 28* (4), 397-412.

Van Driel, J.H., Verloop, N., & De Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching, 35* (6), 673-695.

Verloop, N., Van Driel, J., & Meijer, P. (2001). Teacher knowledge and the knowledge base of teaching. *International Journal of Educational Research, 35* (5), 441-461.

Acknowledgement

This study was funded by the Dutch Organisation for Scientific Research

Author note

Marieke F. van der Schaaf, Department of Educational Sciences, Utrecht University;
Karel M. Stokking, Department of Educational Sciences, Utrecht University;
Nico Verloop, ICLON Graduate School of Education, Leiden University.

Corresponding author: Marieke van der Schaaf, Department of Educational Sciences, Utrecht University, PO BOX 80140, 3508 TC Utrecht, The Netherlands. E-mail: m.vanderschaaf@fss.uu.nl