Setting standards and primary school teachers' experiences of the process

Vanessa Scherman, Lisa Zimmerman, Sarah J. Howie &

Roel Bosker

In South Africa, very few standard-setting exercises are carried out in education and, if they are, teachers are not involved in their execution. As a result, there is no clear understanding of what the standard is and how it was arrived at. This situation is compounded when teachers are held accountable when learners do not meet the prescribed standards. The aim of this paper is to explore how teachers experienced a standard-setting process and if any reflection on teaching and learning practices took place. The research question addressed is: How was the SAMP standard-setting exercise experienced by teachers? Standard setting is an important component in monitoring systems whether national or school-based monitoring systems. The South African Monitoring system for Primary Schools (SAMP) is a school-based monitoring system. The Grade 1 SAMP assessment is administered in English, Afrikaans and Sepedi. As part of SAMP, a standard-setting exercise was undertaken for the Grade 1 assessment to better inform the dissemination of performance results to schools. As part of a bigger research project, a participatory qualitative approach was followed in which 27 participants were asked to partake in a modified Interactive Qualitative Analysis (IQA) process. Several themes were identified by the teachers, namely reflections on practice; emotions before and during the standard setting workshop; benefits of working in a group; and improvements in the process. In this article, these themes are explored and their relevance for standard setting within school contexts discussed.

Vanessa Scherman
University of Pretoria
E-mail: vanessa.scherman@up.ac.za
Tel: 012 420 2498

Sarah J. Howie
University of Pretoria
E-mail: sarah.howie@up.ac.za
Tel: 012 420 4131

Lisa Zimmerman
University of South Africa
E-mail: zimmel@unisa.ac.za
Tel: 012 429 4755

Roel Bosker
University of Groningen
E-mail: r.j.bosker@rug.nl
Tel: +31 50 363 6791

Introduction

Standards in education and the process of generating such standards fulfil critical functions within the education system and are essential in the discussion on what constitutes quality education. At the heart of the educational quality debate is the question of what is being monitored. If levels of achievement are to be monitored, they have to be measured against something, and a standard is the mode of such a measurement. Standards can be described as the degree to which excellence is required for a particular purpose or, more specifically, standards refer to the measure of what is seen as adequate specifically in relation to performance levels (Livingstone, 1988). Put succinctly, standards represent 'how much is good enough' (Bandaranayake, 2008). Performance levels or competence is measured along a continuum, and a standard represents the point which separates competence and incompetence (Bandaranayake, 2008). For monitoring purposes, especially if academic growth of learners is to be tracked, adequate standards and levels of achievement are needed to provide fair and valid information that can be used in schools (Scherman, Howie & Bosker, 2011).

With this in mind, a project was initiated to explore the best manner to identify standards to be used in a school-based monitoring system adapted for South African primary schools, namely the South African Monitoring and Feedback System for Primary Schools (SAMP). The adapted SAMP instruments for Grade 1 were explored in terms of their validity and reliability as well as how the results generated could be used in the classroom. This innovative monitoring system, targeting a variety of school contexts, includes instruments in English, Afrikaans and Sepedi. A baseline and follow-up assessment forms part of the suite of instruments so that academic growth over the year can be tracked.

The main aim of using SAMP as a monitoring system was to inform schools and their own process of self-evaluation. Thus, it was necessary to engage school role players to understand the level descriptors of performance so that the use of the information provided by the monitoring system could be utilised effectively at classroom level (Bandaranayake, 2008). Different standard-setting methods were explored for this purpose (Scherman *et al*., 2011). The main aim of this article is to gain some insight into and describe the experiences of teachers participating in a standard-setting exercise as part of the SAMP. The main research question explored in this paper is: *How was the South African Monitoring Project standard-setting exercise experienced by teachers?*

The standard-setting process and experiences thereof

Standard setting is the process of establishing cut scores (Cizek & Bunch, 2007), while cut scores are thought of as points on a continuum (Thomas & Peng, 2004). Moreover, it is possible to identify more than one cut score on the ability continuum to provide a range of different performance levels so that within each performance level the

knowledge and skills that learners are able to accomplish can be described. If the knowledge and skills can be identified, teachers are better able to target teaching and learning interventions according to the needs of learners at the identified levels.

Setting standards is a complex undertaking (Yorke, 1999) and traditionally can take place using a variety of methods, either in isolation or in conjunction with other methods. Methods such as the Angoff and Nedelsky borderline groups and contrasting groups methods are referred to in the literature (Bandaranayake, 2008; Cizek & Bunch, 2007; Ben-David, 2000; Berk, 1986; Hambleton, 1994). With advances in modern test theory, additional methods are being explored. Item mapping is one such example which utilises Rasch modelling to explore item performance (Wang, 2003). The item difficulties can be presented graphically with all items being ordered from easy to difficult. The aim of item mapping is to locate the item for which there is consensus among the participants that a minimally competent individual would have a 50% chance of answering the item correctly on the achievement continuum (Wang, 2003).

In the past, standard setting has been criticised for being arbitrary (Glass, 1978) based on the assumption that the process involved depends on the subjectivity and imprecision of human judgement (McGinty, 2005). In the years following this initial criticism, researchers have attempted to develop technology for the application of making judgements (Schafer, 2005). With these applications it is possible to set clear and explicit performance standards against which success can be judged (Anderson, 2005). These performance standards provide an indication of what is seen as adequate (Livingstone, 1988). Thus, competence can be measured along a continuum with the standard representing the point which separates competence and incompetence. The specific content of the performance level can then be described in terms of knowledge and skills as tested by the items captured within a performance level (Bandaranayake, 2008; Ben-David, 2000).

When standards are set, participants of the process have to understand that a 'real standard' is a score on an assessment that indicates the level of ability of test-takers in the context of the particular assessment used for setting standards (Klein, 1998). In general, however, participants experience difficulty with the cognitive processes associated with various standard-setting activities, such as interrogating item difficulty and determining what a borderline learner at each performance level would be able to accomplish. As a result, participants tend to rely on alternative strategies such as their experience working with learners and content knowledge (Hein & Skaggs, 2009). Research (Skaggs & Tessema, 2001) has also found that standard-setting participants disagree with the ordering of the items although they do understand the concept of mastery.

To elaborate, there are various factors which may have an influence on how judgements are made, including the workshop process and preconceived ideas about standard setting (Giraud & Impara, 2005). There is also a desire on the part of

the participants to set high standards and to be viewed as doing a good job (McGinty, 2005). To complicate matters, when faced with the array of information provided to aid the process, participants tend to draw on their own experience without necessarily utilising the information available (Ferdous & Plake, 2005). The participants may also arrive at the meetings with very different ideas, and once the standard-setting process is complete participants seem to be reporting more confidence and understanding than they actually have in the process (Skorupski & Hambleton, 2005).

Acknowledging these challenges that participants face in standard setting, the bookmark standard-setting procedure has been developed, which relies on modern test theory. For the purposes of this research, the bookmark procedure was deemed to be the most appropriate procedure to implement as subjective human judgement is combined with the empirical exploration of item characteristics such as item difficulty and associated learner ability. It is described further below.

The bookmark standard-setting procedure

The bookmark procedure was described by Lewis, Mitzel and Green in 1996. This method uses an Item Response Theory (IRT) or a Rasch item-mapping procedure to order items by difficulty, thereby simplifying the cognitive task of participants as they do not have to decide on the difficulty of the item themselves (Buckendahl, Smith, Impara & Plake, 2002). In this approach, items are ordered from easiest to most difficult (on a common scale or continuum) and are bound in a book format. Items are, thus, mapped to locations on this difficulty scale so that individuals with scale scores near the location of items can be inferred to have the knowledge, skills and abilities which are required to respond successfully to the items.

As part of the bookmark method, participants are then required to place a bookmark between the most difficult item and what a minimally competent or borderline learner would be expected to answer correctly. Each participant is requested to evaluate the item in terms of whether a minimally competent individual would answer the item correctly. A difficulty parameter of 0.5 is used as a guideline for participants so that they can think in terms of whether the learner has a 50% chance of answering the item correctly (Lissitz & Kroopnick, 2007; Schagen & Bradshaw, 2003). If participants agree that a minimally competent individual would have more than a 50% chance of answering correctly, the next item would be the position of the bookmark. The next item would typically be harder than the previous one. In an iterative process, the participants are given an opportunity to discuss their evaluations as well as make revisions based on the discussions in the group (Lissitz & Kroopnick, 2007; Schagen & Bradshaw, 2003).

The bookmark method is widely used within the field of education due to its logical appeal and practicality. The most important consideration for selecting the bookmark method is that the literature suggests that participants in the process understand and have confidence in the procedures and activities undertaken (Buckendahl *et al.,*

2002; Dawber *et al.*, 2002). A criticism against the bookmark method is that arbitrary values are used to establish the point along the ability continuum that is used to rank order the items (Buckendahl *et al.*, 2002).

For the standard-setting procedure for the SAMP implementation at Grade 1, four levels were decided upon: *below basic*; *basic*; *proficient*; and *advanced* (Buckendahl *et al.*, 2002). The standard-setting exercise took place over the course of four afternoons. The bookmark procedure provided a complete set of activities designed to yield cut scores in a methodologically rigorous approach and was relatively easy to implement (Cizek & Bunch, 2007).

Research methodology

Exploration of the standard-setting workshop for SAMP by means of qualitative research in the form of a modified Interactive Qualitative Analysis (IQA) allowed for access to teachers' experiences thereof (Northcutt & McCoy, 2004). This exploration was conducted through a mutual process of meaning sharing and interpretation between the researcher and participants (Berg, 1998). Many qualitative studies espouse a goal to capture the meaning of a phenomenon from the research participant's point of view. However, according to Northcutt and McCoy (2004), there may be grounds to question this by explaining that, in many qualitative studies, the participant and the researcher are separated by means of the institution of a power-based hierarchy which favours the researcher's expert status as the role-player with the sole ability to generate and analyse data. The ability of participants to add analytical value to the research is largely ignored, which leads to the questioning of the objectivity and accuracy of the research findings. In contrast, the modified implementation of IQA used in this study allowed participants to add analytical value to the research and, in so doing, contributed to the overall rigour of the study.

Participants

Learners and teachers in 43 schools participated in the SAMP project. These schools were purposively selected based on geographic area as well as language of learning, namely English, Afrikaans and Sepedi. All of the schools were invited to participate in the standard-setting workshop. Twenty-seven teachers responded to the invitation and were predominantly from the Afrikaans (language of instruction) schools. On average the participants had been teaching for 13 years and had been teaching at their current school for 8 years.

Data collection via selected processes from the Interactive Qualitative Analysis method

The IQA method, as described by Northcutt and McCoy (2004), when followed in its entirety, is a social constructionist approach to data collection and analysis. It addresses power relations between researcher and participant by encouraging greater participation of participants in the (1) *elicitation of themes* via collaborative axial coding and the creation of theory; and in (2) *data analysis* via the use of Affinity Relationship Tables (ARTs). In this way participants can record their view of the possible relationships amongst themes, leading to the creation of a systemic interrelationship diagram. This analysis method ensures that participants are able to add analytical value to the research.

Typically, IQA involves a focus group that generates qualitative content in the form of codes, which the group then arrange into categories of meaning, providing a central theme that binds each of the codes generated together into a category. Each individual will provide indications of the relationships amongst these categories via a cause–effect analysis. Ultimately, after a quantitative analysis of the relative frequency of each possible relationship given by all the members of the group, a mind map of the group's system of meaning will be drawn up using this quantitative information as a guide. For the purposes of this research, and due to time constraints, only the first step in the IQA process was implemented, namely involving the participants in the generation of their own data and elicitation of themes during a focus group. This strategy was deemed most appropriate for a standard-setting exercise in which the participation of the teachers and their own meanings was considered to be paramount.

This modified data collection process was initiated as part of the standard-setting workshop which took place over four afternoons from Tuesday to Friday. In each workshop session on each afternoon teachers were given feedback based on the cut scores (bookmarks) they had identified the previous day. Discussions as to why cut scores were collected were undertaken. After the discussion on the final cut scores teachers were asked to reflect on the activities for the week and to write down their ideas on note cards they had been provided with. Practically, this meant that, in order to provide some form of guidance to the participants, the researchers asked them to think about the standard-setting process followed during the week. They were specifically asked to reflect on what they did, what they discussed, how they felt before and during the standard-setting process, what worked well and what they thought could be changed.

Once teachers had captured their ideas on the note cards they were requested to stick these on the wall. This resulted in what can be described as a collage of teacher participant generated codes on the wall. Teachers were then provided with the opportunity to read each other's ideas from the note cards and they could also

clarify any meanings. Then teachers grouped the ideas together based on what they thought were common themes; an activity traditionally undertaken by the researcher following qualitative coding. These discussions were captured and then transcribed.

Scientific rigour

To achieve rigour in qualitative research, Northcutt and McCoy (2004) recommend data collection and analysis methods that are public and non-idiosyncratic and that do not depend on the nature of the codes themselves. In order to attain this, the participants were required to create themes. It also served as a member-checking strategy because it was used as a means of consequent contact with the participants in the study to check whether they were in agreement. Member checking, as such, can also be referred to as 'respondent validation' *(*Taylor, 2001).

Ethical considerations

Before the workshop started participants were asked to complete informed consent forms explaining the process of the research and the purposes for which it would be used. Participants were briefed on the workshop process and were informed that they could withdraw from the research at any time. Bearing in mind the limitations of focus group research with regard to anonymity, confidentiality for the overall project and reporting was confirmed.

Limitations of the process

The standard-setting workshop took place over a period of four afternoons. The modified IQA process took place on the final day and the teachers appeared to be experiencing fatigue. Participant attrition was also problematic as three teachers did not return to the workshop after the first day. Another aspect was the level of uncertainty of the teachers. This was something new to them and it took some time to settle into the standard-setting process.

Findings: Teachers' self-generated themes

As indicated teachers were asked to reflect on the process of the standard-setting workshop and, in order to provide some structure, they were specifically asked to think about what they did, what they discussed, how they felt before and during the standard-setting process, what worked well, and what they thought could be changed. Five themes were identified by the teachers, namely *reflections on practice*; *emotions before the standard-setting workshop*; *emotions during the standard-setting workshop; benefits of group work;* and *improvements in the process*.

Reflections on practice

Teachers tended to think about the manner in which they were assessing at school and that perhaps they should think about their assessment practice. For example, teachers generated the following codes as part of this theme:

> *It made me think about my way of assessing;*
> *That I must put more thought into it;*
> *We shared interesting ideas and the way of doing things in classes;*
> *I gained insight on the procedures at my own school and other schools.*

Teachers also started to reflect on what their interpretation of a standard is and the realisation that everyone does not necessarily have the same understanding. Examples of this are:

> *Dit was leersaam om te besef dat ander se persepsies oor assesseringstandaarde nie noodwendig dieselfde as myne of die skool s'n is nie (It was informative to realise that others' perceptions about assessing standards are not necessarily the same as mine [or the school's])*

> *I've learned of how different each school or level of the work is and the process would work if there is at least an improvement from basic to the follow up.*

Perhaps one of the most insightful reflections captured by one of the teachers was: *I found it interesting to see how my thoughts changed after group discussions.* This may give an indication of the level of engagement of the teachers, that they thought about the rationale of other teachers for selecting a cut score, put their ideas forward and that consensus was reached.

Emotions before and during the workshop

Teachers seemed to feel some trepidation and uncertainty about the standard-setting process and what it would entail:

> *No idea what the process is all about; Never thought that standard setting is such a big thing.*

> *Initially I was negative regarding the 4 days due to lack of information what the 4 days would entail.*

Many of the teachers were negative at the beginning of the workshop, thinking that this would be something that they would find difficult to undertake:

> *It was going to be difficult for me, we are going to set questions.*

> *Initially I was nervous and scared for some reason thinking I won't know enough to participate but I then realised I would also be learning about certain teaching abilities that I wasn't aware of and gaining more knowledge where teaching is concerned.*

Other teachers were unsure of their feelings regarding the standard-setting process

> *At first I wasn't sure how I felt about this and what to expect.*

Many of these negative perceptions were changed during the standard-setting workshops as participants indicated that they

> *... felt free, I found things easy and simple to read and work with them.*

> *No restriction, all opinions welcomed and encouraged, not bored, programme clear and we knew where we were going (goals clear).*

It seemed that teachers had a positive view of the standard-setting workshop and felt safe to participate and share their ideas:

> *Opgewonde, positief om te sien hoe en watter prosesse gebruik word om die standaarde te bepaal. (It is exciting to see how and what processes are used to determine the standards.)*

Benefits of working in the group

A recurrent theme was that, by working in groups, teachers had the opportunity to share ideas, listen to one another, and have the opportunity to talk through their thinking. Examples of this include:

> *Groepe is 'n baie goeie idee – kry nuwe insigte en perspektief van kollegas van ander skole. (Groups are a very good idea – getting new insights and perspectives from colleagues from other schools.)*

> *In the end it was really interesting to see the different views of other educators; setting the standards is a GREAT idea.*

Participants felt that the *process was well thought through and everything ran smoothly,* that the *discussion and participation by all was informative and enlightening* and that *the information provided was well presented and interesting*, which could be discussed further within the groups.

Improvements in the process

While all the schools participating in the SAMP were invited to the standard-setting workshop, the majority of the participating teachers came from Afrikaans schools. Teachers from three English schools and one Sepedi school also participated in the workshop. For this reason the participants suggested that a *wider variety of schools should be present so as to standardise the results.* One participant specifically suggested *3 inner city, 3 suburban (Eng(lish) and Afrik(aans)) and 3 township* schools. Some teachers also felt that the standard-setting process was drawn out and that everything could be managed in two days by lengthening the sessions on the days:

> *Ek sou die proses korter wou maak net 3 of 2 dae met bv. bietjie langer sessies. (I would have wanted to shorten the process to three or two days with for instance slightly longer sessions.)*

Finally, some teachers felt that the groups should be moved around between rounds or that the groups should be made bigger:

Mixing the groups on round 2 and 3 or making bigger groups could be more
beneficial; after session 1 change groups to get other people's views.

Relationships between the themes generated and literature

The literature suggests a level of uncertainty associated with the standard-setting process (Giraud & Impara, 2005; McGinty, 2005). However, during the standard-setting procedure participants realised that they had something to offer, an expertise of their own. This is normally done by means of drawing on their own experience (Ferdous & Plake, 2005), a process facilitated by the modified IQA. They obtain the insight that they have something to offer by means of reflecting on their practice. By working in groups they have the opportunity to voice their opinions, listen to others and perhaps revise their previous thinking. This all contributes to the positive emotions that they are already experiencing. Based on their reflections they are also able to make valuable suggestions for future standard-setting workshops.

As stated above, the next step in the IQA process, (namely *data analysis* via the use of Affinity Relationship Tables [ARTs] so that participants can record their view of the possible relationships amongst themes) was not feasible due to time constraints for this study. Instead, as such, a researcher-generated understanding of the possible impact of the process of the workshop standard-setting exercise is provided in figure 1 as a discussion point for further research in this area.
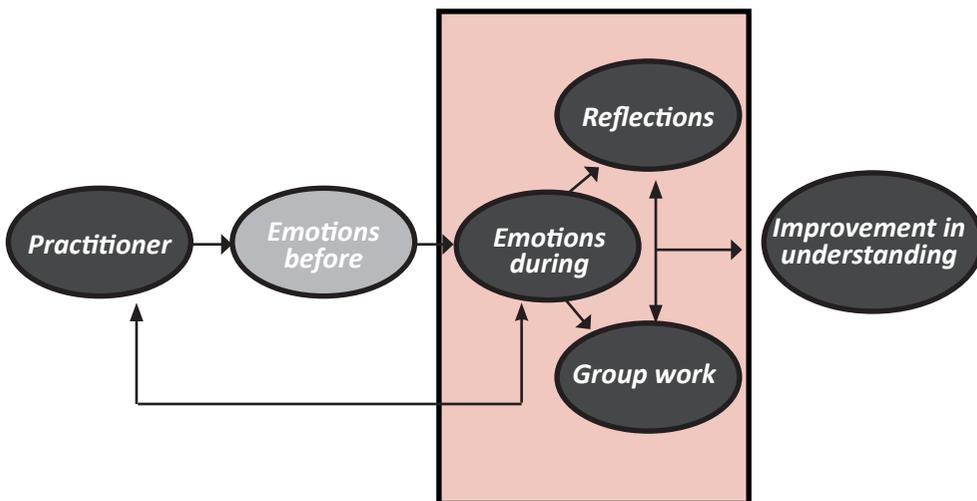

Figure 1: Visual representation of the themes.

Conclusion

While the literature suggests that identifying the borderline candidates at each of the levels can be challenging initially, this research revealed that it is possible for teachers to adapt to the process and contribute meaningfully. The participants in this

study tended to arrive at the workshop with very different ideas as to what would take place. However, with support and collaborative discussions participants gained confidence and deepened their understanding not only of the standard-setting process, but also of possible teaching practice applications. Participants enjoyed the interaction with other teachers, the discussions often changing preconceived ideas in light of new arguments. In the South African context, teachers would have appreciated a more diverse participant group. Furthermore, in a context where teachers, to a large extent, feel marginalised, participatory events such as these may instil a sense of value in the contributions they make. In the words of one participant '*thank you for treating us with respect, as if we have something to add'.*

As is indicated in the literature, the bookmark method is understandable and practical to implement. However, initially, teachers disagreed with ordering of the items based on the empirical difficulty of the item as it seemed counter intuitive. Teachers also seemed to initially rely on their experiences with learners as opposed to taking the item difficulty into account. This, however, was addressed by additional discussions of what the item difficulty means. Teachers also tended to assist one another to obtain a common understanding of the procedure.

While this paper focuses on the experiences of teachers it is important to also consider future applications for standard-setting exercises. One of the main limitations of the process was that teachers in a sense were self-selected. The participants were predominantly from a particular school context and this may have had an influence on the initial variation of cut scores. For future sessions individual schools will need to be approached based on maximum variation so that the performance levels may be applicable across the school contexts. Furthermore, as two assessment instruments are involved, the way in which the performance levels are described has to be carefully considered so that academic growth can be plotted adequately.

Acknowledgements

References

Anderson J 2005. *Accountability in education.* Paris: UNESCO.

Bandaranayake RC 2008. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher,* 30: 836-845.

Ben-David MF 2000. AMEE Guide No. 18: Standard setting in pupil assessment. *Medical Teacher,* 22: 120-130.

Berg BL 1998. *Qualitative research methods in the social sciences* (3rd ed.). Boston: Allyn & Bacon.

Berk RA 1986. A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research,* 56: 137-172.

Buckendahl CW, Smith RW, Impara JC & Plake BS 2002. A comparison of Angoff and

Bookmark standard setting methods. *Journal of Educational Measurement*, 39: 253-263.Cizek GJ & Bunch MB 2007. *Standard setting: A guide to establish and evaluate performance standards on tests.* London: Sage Publications.

Dawber T, Lewis DM & Rogers WT 2002. The cognitive experience of Book-mark standard setting participants. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Ferdous AA & Plake BS 2005. Understanding the factors that influence decisions of panelists in a standard setting study. *Applied Measurement in Education*, 18: 257-267.

Glass GV 1978. Standards and criteria. *Journal of Educational Measurement*, 15: 237-261.

Giraud G & Impara JC 2005. Making the cut: The cut score setting process in a public school district. *Applied Measurement in Education*, 18: 289-312.

Hambleton RK 1994. Standard setting in criterion referenced testing. In T Husen & TN Postlethwaite (eds), *The International Encyclopaedia of Education*. New York: Pergamon Press.

Hein SF & Skaggs GE 2009. A qualitative investigation of panelists' experiences of standard setting using two variations of the bookmark method. *Applied Measurement in Education*, 22: 207-228.

Klein SP 1998. Standards for teacher tests. *Journal of Personnel Evaluation in Education,* 12: 123-138.

Lissitz RW & Kroopnick MH 2007. An adaptive procedure for standard setting. Retrieved on 12 July 2009 from: http://www.education.umd.edu/EDMS/MARCES/presentation/Lissitz_KroopnickFNLNCME07.pdf.

Livingston ID 1988. Standards, national monitoring. In T Husen & TN Postlethwaite (eds), *The International Encyclopaedia of Education*. New York: Pergamon Press.

McGinty D 2005. Illuminating the 'black box' of standard setting: An exploratory qualitative study. *Applied Measurement in Education*, 18: 269-287.

Northcutt N & McCoy D 2004. *Interactive qualitative analysis*. Thousand Oaks: Sage.

Schafer WD 2005. Criteria for standard setting from the sponsor's perspective. *Applied Measurement in Education*, 18: 6-81.

Schagen I & Bradshaw J 2003. Modelling item difficulty for bookmark standard setting. Paper presented at the British Educational Research Association Annual Conference September, Edinburgh, United Kingdom.

Scherman V, Howie SJ & Bosker 2011. Constructing benchmarks for monitoring purposes: Evidence from South Africa*. Educational Research and Evaluation*, 17: 511-525.

Skaggs G & Tessema A 2001. Item-disordinality with the Bookmark standard setting procedure. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Skorupski WP & Hambleton RK 2005. What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education*, 18: 233-256.

Taylor S 2001. Evaluating and applying discourse analytic research. In M Wetherell, S Taylor & SJ Yates (eds), *Discourse as data: A guide for analysis.* London: Sage Publications.

Thomas S & Peng WJ 2004. The use of educational standards and benchmarks in indicator publications. *European Education Research Journal,* 3: 177-212.

Yorke M 1999. Benchmarking academic standards in the UK. *Tertiary Education and Management,* 5: 81-96.

Wang N 2003. Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement,* 40: 231-253.