Art. #1974, 9 pages, https://doi.org/10.15700/saje.v41n4a1974

Applicability of two standard setting methods for enhancing the reporting of assessment results within the South African education context

Qetelo Moloi and Anil Kanjee
Department of Primary Education, Tshwane University of Technology, Soshanguve, South Africa
kanjeea@tut.ac.za

The study reported on here contributes to the growing body of knowledge on the use of standard setting methods for improving the reporting and utility value of assessment results in South Africa as well as for addressing the conceptual shortcomings of the Curriculum and Assessment Policy Statement (CAPS) reporting framework. Using data from the "verification" version of the Annual National Assessments (ANAs), we explored relevant technical and conceptual factors to consider for the application of standard setting methods. Two sets of panellists were trained to generate cut scores for Grade 6 mathematics and English First Additional Language (FAL), one using the Angoff method and the other the Objective Standard Setting (OSS) method. The findings indicate that the 2 methods generated different sets of cut scores across the performance levels for both subjects. While these cut scores had significant implications for the percentage of learners classified at each performance level, they were consistent with findings from other studies. We also identified 4 key factors to address when undertaking standard setting exercises: engagement with test content, resource requirements, requisite expertise and software, and collective accountability. We conclude that standard setting approaches should be the preferred option to the CAPS reporting framework when reporting assessment results in South Africa. More importantly, the decision on the most appropriate method for the South African context depends largely on the extent to which the 4 key factors identified can be addressed.

Keywords: Angoff method; OSS method; performance standards; standard setting

Introduction

The implementation of the new systemic evaluation model by the Department of Basic Education (DBE) is intended to address several limitations of the Annual National Assessment (ANA), the most important of which is to enhance the use of the assessment results to improve teaching and learning in schools (DBE, Republic of South Africa [RSA], 2017). Currently, the results of learner performance, from both large-scale assessments and public examinations in South Africa are reported according to the framework specified in the CAPS document (DBE, 2011). The CAPS reporting framework lists "Rating codes" that indicate the level of performance based on percent-correct responses obtained by candidates. These are presented on a seven-level hierarchical scale, each level associated with a specific percentage range and label and are categorised as: 7: Outstanding Achievement (80%–100%), 6: Meritorious Achievement (70%–79%), 5: Substantial Achievement (60%–69%), 4: Adequate Achievement (50%–59%), 3: Moderate Achievement (40%–49%), 2: Elementary Achievement (30%–39%) and 1: Not Achieved (0%–29%) (DBE, 2011:302).

However, several features of the CAPS reporting format compromise effective and meaningful reporting. Firstly, the rating codes make no provision for detailed reporting on what learners at each level can or cannot do, nor the knowledge and skills that learners may or may not command. The absence of descriptive detail limits the use of the results to identify learner weaknesses and strengths for appropriate intervention. For example, a score of 66% provides no information on the specific learning needs of the learners who obtained this score, and thus, that which ought to be done to address these needs. Secondly, the same rating codes are applied to all subjects, implying an equal weighting across different subjects, notwithstanding the possible variations in cognitive demands across the subjects. Thirdly, the use of seven hierarchical levels may suffer decreased accuracy in reported performance at each level since, from most tests or examinations, the items would either be clustered in fewer levels or would be stretched too thinly across all levels (Sondergeld, Stone & Kruse, 2018; Zieky & Perie, 2006). Fourthly, the CAPS score bands are based on an incorrect assumption that cognitive demand across the different intervals are homogeneous (Bond & Fox, 2007). For instance, the interval of 15–25% is 10% and equal to 85–95%, but the cognitive demands at the respective intervals are not equal. A questionable assumption in using the CAPS framework is that improvement of performance from 15% to 25% is equivalent to improvement from 85% to 95%.

The use of standard setting (SS) approaches offers a viable option to addressing these limitations (Baird, Isaacs, Opposs & Gray, 2018; Bejar, 2008; Griffin & Nix, 1990; Haertel, 2005). SS approaches provide a novel way of organising, processing and reporting examination and assessment data in more user-friendly formats. Similar to many other countries, there have been several initiatives to apply SS within the schooling and higher education sector in South Africa (Kanjee, Claassen, Makgamatha & Diedricks, 2004; Kanjee & Moloi, 2016; Moloi & Kanjee, 2018; Pitoniak & Yeld, 2013; Scherman, Zimmerman, Howie & Bosker, 2014). While different SS approaches have been used in these studies, there has been no attempt to investigate whether any of the different approaches would be more appropriate for use within the South African education context.

Standard Setting in South Africa

Initiatives to introduce SS to report assessment results in South Africa have revealed several possibilities and setbacks. Kanjee et al. (2004) applied the Angoff method to report results of a Grade 9 national assessment survey of English, mathematics and science. The assessment was conducted using matrix sampling approaches, and analysis was undertaken using Item Response Theory (IRT) to determine learner performance. While the results provided detailed information that policymakers, district officials and teachers regarded as valuable for use in developing interventions, the authors noted that the SS exercise proved extremely costly. This was due to the time required to process the large number of items for each subject area, the costs of specialised IRT software, as well as the costs of involving panel members, and for undertaking analysis that required highly technical expertise and experience.

In their study, Kanjee and Moloi (2016) compared cut scores obtained from the Angoff SS method to cut scores calculated according to the CAPS reporting scale (DBE, 2011). The findings showed that the cut scores from the two methods categorised learners differently, with the CAPS cut scores classifying a higher percentage of learners at the lowest levels of performance compared to the standards-based cut scores. However, the authors noted that these findings could not be conclusive because there was no way of verifying whether the discrepancies were real, or were ascribable to the specific reporting method used. Pitoniak and Yeld (2013) also used the Angoff SS method to report results of the National Benchmark Tests (NBT), an assessment used by some South African universities as part of their admissions process. Their study highlighted some of the constraints that the unique apartheid legacy placed on the feasibility of using SS methods in the South African context. Pitoniak and Yeld (2013) report that the panellists' deep sense of mistrust had an impact on both the processes and the resulting standards. Specifically, the authors noted that panellists were of the view that standards-based reporting would perpetuate race-based inequalities in education, and further disadvantage students from poor and marginalised backgrounds.

Scherman et al. (2014) used the Bookmark method (Näsström & Nyström, 2008) to investigate teachers' experiences from participation in a SS process that involved learner tests that were administered in three regional languages. The authors reported that the majority of the panellists comprised teachers from one language group that also represented high performing and well-resourced schools. Commenting on the skewed language and social class representation of the panellists, Scherman et al. (2014) noted that this could have influenced the determination of cut-scores, and cautioned that participants in any SS process must be representative of the South African schooling context for standards-based reporting in order to be valid and reliable.

The foregoing studies demonstrated the need to account for specific contextual factors when using appropriate SS approaches for reporting assessment results, while also highlighting a range of conceptual and technical factors that ought to be considered. Consequently, a concern arose whether specific SS methods would be more suitable and acceptable to the South African context. In addressing this concern, we report the results of a comparison between two established SS approaches; the Angoff (Angoff, 1971) and the OSS method (Stone, 2001).

This article contributes to research-based techniques and procedures for determining cut scores in the use of SS processes within contexts that are marked by limited resources and capacity, are traditionally examination-dominated and are transforming to using low-stakes large-scale assessments for purposes that include diagnosis of what learners know and can do. The focus of the article is on researching the contextual and technical factors to consider in the development and deployment of the Angoff and OSS methods within the specific context of the education in South Africa rather than on reporting on a particular study or generalising findings from the data used.

We identified the Angoff method, given that it is one of the most widely used methods by examination and licensure bodies (Hambleton & Pitoniak, 2006; Näsström & Nyström, 2008), while the OSS method was identified as having been developed primarily to address the weaknesses of the Angoff method (Stone, 2001). In the next section we present the conceptual framework applied, followed by the research questions addressed, and the methodology used. Next, the findings and discussions are presented, while we conclude by listing several limitations and options for additional research studies.

Conceptual Framework

Essential to any SS process is the requirement to distinguish between content and performance standards (Cizek & Bunch, 2007). Content standards answer the "what" question, that is, what do learners need to know and be able to do? In most education systems, this content is specified in the curriculum documents. Performance standards, on the other hand, answer the "how much" question, that is, they measure how much progress learners have made in what they know and can do.

The determination of "performance standards", however, requires SS, a deliberate process that involves participation and approval of policymakers as well as engagement and judgements from subject experts. While this process focuses largely on the technical aspects, it is also steeped in diverse conceptual and theoretical understandings on the use of data for decision-making, as well as the specific policy and practice context of teaching and learning that define different education systems. More importantly, the growing body of literature on the merits and demerits of different methods of SS have highlighted a range of findings that have led to several researchers concluding that there is no best method of setting performance standards, where the reasoning is forwarded that the best method is one that fits the purpose of the user (Näsström & Nyström, 2008; Tiratira, 2009).

The SS process begins by defining and labelling Performance Levels (PLs), followed by specifying Performance Level Descriptors (PLDs), establishing cut scores and concluding with the approval of the standards and cut scores by the relevant authorities (Haertel, 2005; Hambleton & Pitoniak, 2006). PLs refer to the general policy statements that indicate the official position of the relevant authorities on the desirable number and labels of categories to be used in classifying learners according to their knowledge and skills in a particular subject (Zieky & Perie, 2006). PLDs are detailed descriptions of "the knowledge, skills, and abilities to be demonstrated by students who have achieved a particular performance level within a particular subject area" (Zieky & Perie, 2006:4). The PLs and PLDs for mathematics and FAL that were used in this study were developed by respective teams of teachers and subject specialists of the DBE. Four levels were developed, namely: PL1 – Not Achieved (NA); PL2 – Partially Achieved (PA); PL3 – Achieved (AC); and PL4 – Advance (AD). Additional technical details on the process and PLDs are reported in Kanjee and Moloi (2016). Within this conceptual framework, the following research questions were addressed in this study:

1) How do cut scores generated through OSS and Angoff procedures of SS compare?
2) What is the impact of the cut scores from each method on categorising learners at the different PLs?
3) Which method would be most appropriate for the South African context?

Methodology

In this section we report on the participants involved in the SS process, the data used, the specific steps applied for each SS method, and the analysis conducted.

Participants

Panel members were selected from teachers who were appointed by the DBE to develop items for the ANA tests. Only teachers who were subject area specialists in Grade 6 mathematics and English, and who had extensive teaching experience as well as expertise in item writing and test development were selected. The final group of panellists were representative of teachers from the different quintile school categories. Table 1 lists the number of panellists participating in each SS method and subject.

Table 1 Panellists involved in the SS processes

	Number of panellists		
Subject	Angoff	OSS	Total
FAL	5	5	10
Maths	8	7	15
Total	13	12	25

Data

The quantitative data used in the study included learner test scores and ratings of test items by panellists, while the qualitative data were obtained as feedback from the panellists before, during, and after the SS processes. It was crucial that the learner test scores were based on reliable and valid datasets that were applied in the South African context. Thus, test scores were obtained from the verification version of the 2013 ANAs administered to a stratified random sample of Grade 6 learners in mathematics (*n* = 8,131) and English FAL (*n* = 6,106) (DBE, RSA, 2013). The reliability indices for both tests were 0.90. The verification version of the ANAs were administered and scored by an independent, external agency under controlled conditions, similar to the high-stakes Grade 12 examinations and also similar to how future systemic evaluations will be administered and marked (Mweli, 2018). Moreover, it is important to note that we are not reporting on the 2013 ANA results, but are demonstrating the deployment of SS techniques on data that resembles, in both collection and validation processes, future datasets for SS. Details about the test content, sampling, administration, and marking of the ANA data can be accessed from DBE, RSA (2013).

Procedure

Each SS session began with the training of panel members followed by the process for determining cut scores. The training focused on equipping panel members with the techniques and processes that they had to apply in determining cut scores. Table 2 presents the content covered at each training session.

Table 2 Key training focus areas for the Angoff and OSS methods

Angoff methods	OSS method
Overview: The use of SS in education	
Definition of "minimally competent" or "border-line" learners	Definition of "essential" and "non-essential" test items
Taking the test for familiarisation with the items	
Overview: Process for rating of test items Practice exercise: Rating of items (3 iterations) and calculation of cut scores	Overview: Process for categorisation of test items Practice exercise: Categorising of items into "essential" and "non-essential" groups
Evaluation of process	

Angoff SS method

The Angoff method involves panellists estimating the competence of a minimally competent hypothetical borderline learner for each item in a given test (Hambleton & Pitoniak, 2006). A hypothetical borderline learner is defined as one who functions at the interface of two adjacent PLs (Zieky & Perie, 2006). The panellists were introduced to the theory and techniques of the Angoff method, and given an opportunity to answer the test items in order to gain a deeper understanding of the responses that were expected from leaners. In addition, a practice run was undertaken to ensure that the panellists fully understood the SS process, during which all questions and uncertainties were addressed.

In the first round, panellists worked individually to rate items at each PL and recorded their ratings on a specially designed form. The ratings were captured and results of calculations regarding the mean ratings and range per item across raters were provided to panellists to interrogate and discuss in their subject groups. In the second round, panellists repeated the item rating process, but this time, with the benefit of inputs received from the group discussions. The new set of ratings were captured, and panellists were again provided with feedback, but this time they were also provided with item difficulty values for each item. Final determination of Angoff cut scores in each subject involved averaging ratings from the third round over items and panellists for each PL. The cut score for each PL was then calculated as the average rating expressed as a percentage of the maximum possible rating for each PL. These scores were presented to panellists for final review, and through a process of discussion, adopted as the final cut scores to be submitted to the DBE for approval.

OSS method

The distinguishing feature of the OSS method is the involvement of panels in identifying the essential items from a pool of items presented in the administered examination (Stone, 2001). Essential items are defined as items assessing the most important and critical content knowledge and skills, which learners must answer correctly to be considered to be functioning at a particular PL for the given grade (Stone, 2001). The panellists were introduced to the theory and techniques of the OSS method (Stone, 2001), focusing on ensuring that they were particularly able to distinguish between "essential" and "non-essential" test items at each PL (Stone, 2001:199).

Panellists were provided with a prepared rating sheet that listed the numbers of the items in exactly the same order in which they were numbered in the test and included three columns labelled "PL2", "PL3", and "PL4", respectively. Starting from the first item in the test, and using their knowledge of the curriculum expectations for the subject and grade, each panel member was required to categorise each item in the test as being "essential" under the appropriate PL. For instance, if Item 5 was considered as "essential" for a learner in Grade 6 to answer correctly at the "PL2", panellists simply placed a tick next to Item 5 under "PL2."

A practice run of the actual process was also undertaken to ensure that panellists fully understood the SS process, during which all questions and uncertainties were addressed. Thereafter, panel members worked independently to classify each item in a subject under each PL. Before each panellist's completed form was accepted, it was checked to ensure that (i) all the test items had been classified and that (ii) each item was classified exclusively under one PL. For each subject, all the information was captured onto one comprehensive spreadsheet, where each tick made by a panellist for each item was replaced with a Rasch value for the average difficulty (in logits) of the item, calculated using the Winsteps programme (Linacre, 2014). Final analysis involved averaging ratings, expressed in logits, over items and panellists under each PL in each subject. Cut scores (in logits) were then calculated at each PL using the formula in Equation 1 (Stone, 2001:192–193).

Cut score = "Criterion Point" + Mastery Level ± Confidence Level *(1)*

where "Criterion Point" refers to the mean difficulty of "essential items", "Mastery Level" refers to comprehensive competence required to achieve the relevant PL, and "Confidence Level" refers to the acceptable error band around the cut score. For this study, a "Mastery Level" of 70% (0.85 logits) and confidence level of 95% (1.96 logits) was adopted. The "Confidence Level" adopted for this study was 95% (1.96 logits).

Analysis

Given the categorical nature of the data and the fact that assumptions of normal distribution and homogeneity could not be made, non-parametric methods of analysis were employed. To compare the resulting cut scores across the PL categories in each subject, the rank-based Mann Whitney *U* test (two-tailed) was used to determine whether the observed differences generated through the two methods were statistically significant, or were merely due to ubiquitous chance errors (Monahan & Ankenmann, 2005).

To determine the magnitude of the differences in the cut scores resulting from the two SS methods, the effect size (*d*) was also calculated (Cohen, 1992). As a benchmark, Cohen (1992) suggests that values up to $d = 0.1$ indicate differences of small magnitude, up to $d = 0.3$ differences of medium magnitude, and $d = 0.5$ and above indicate differences of large magnitude in the quantities that are being compared. For comparing the impact that the adoption of the cut scores would have in terms of the percentage of learners categorised at each PL, the Chi-square test of independence was used (Fritz, Morris & Richler, 2012).

Limitations of the Study

In this study, we used two sets of panels who worked separately to determine performance standards using either the Angoff or OSS method. Although thorough panellist training was provided, we had no way of establishing whether one group understood both the concept and the associated techniques the same way as the other group. In addition, while panellists in both studies were representative of schools from the different quintile categories, it was not possible to determine the specific level of expertise and content knowledge of each panel member.

Results and Discussions

The results and discussion for each research question are presented below.

Question 1: How do Cut Scores Generated through OSS and Angoff Procedures of SS Compare?

A summary of the findings on the mathematics and English cut scores is presented in Tables 3 and 4, respectively. At the PA level we found a non-significant effect indicating that there were no differences in the cut scores between the two SS methods. The mean ranks for the Angoff and OSS methods were 7.3 and 8.9, respectively, where $U = 22$, $z = -0.78$, $p > 0.05$.

At the AC level, cut scores for the Angoff method were significantly higher and substantively larger in magnitude ($d = 0.6$). The mean ranks for Angoff and OSS methods were 10.3 and 5.4, respectively ($U = 10$, $z = -2.33$, $p < 0.05$, $d = 0.6$). Similarly, at the AD level, cut scores for the Angoff method were significantly higher and substantively larger in magnitude. The mean ranks for the Angoff and OSS methods were 4.3 and 11.5, respectively ($U = 0$, $z = -3.6$, $p < 0.05$, $d = 0.9$).

Table 3 Mathematics cut scores from two SS methods

SS	Cut scores at each PL		
method	PA (%)	AC (%)	AD (%)
Angoff	26.4	56.7	72.9
OSS	25.9	47.5	59.9
Difference	Not statistically significant	Statistically significant	Statistically significant
Effect size	0.2	0.6	0.9

At the PA level for English, the cut scores for the OSS method were significantly higher and substantively larger in magnitude ($d = 0.8$). The mean ranks for the OSS and Angoff methods were 8.0 and 3.0, respectively ($U = 0$, $z = -2.38$, $p < 0.05$). However, at both the AC and AD levels, the cut scores for the Angoff method were significantly higher, and substantively larger in magnitude ($d = 0.8$). At the AC level, the mean ranks for the OSS and Angoff methods were 3.0 and 8.3, respectively ($U = 0$, $z = -2.38$, $p < 0.05$), while at the AD level the mean ranks for the OSS and Angoff methods were 3.0 and 8.0, respectively ($U = 0$, $z = -2.38$, $p < 0.05$). The common trends across both subject areas were that higher cut scores were generated from the Angoff method at the AC and AD levels.

Table 4 English cut scores from two SS methods

	Cut scores marking different PLs		
SS method	PA (%)	AC (%)	AD (%)
Angoff method	33.2	64.6	85.2
OSS method	39.7	50.5	62.2
Difference	Statistically significant	Statistically significant	Statistically significant
Effect size	0.8	0.8	0.8

Question 2: What is the Impact of the Cut Scores from Each Method on Categorising Learners at the Different PLs?

Figures 1 and 2 show the percentages of learners who would be categorised at different PLs in English and mathematics, respectively, using cut scores generated from the Angoff and OSS methods.

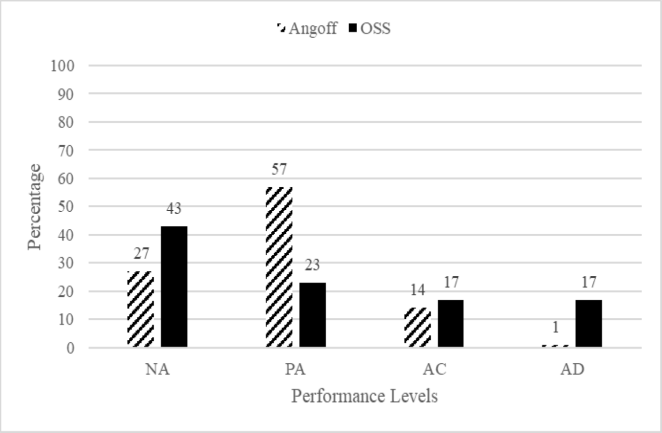

Figure 1 Percentage of learners categorised by the OSS and Angoff methods in each PL for English

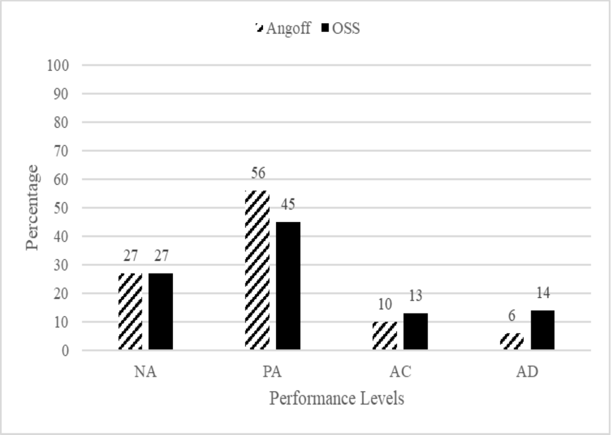

Figure 2 Percentage of learners categorised by the OSS and Angoff methods in each PL for mathematics

A Chi-square test of independence indicates a significant relation of dependence for English ($X^2(3, n = 6{,}106) = 1986.62, p < .05$) with the OSS method categorising a substantially higher percentage of learners at the NA (16%) and AD (16%) levels, the Angoff method categorising substantially higher percentages of learners at the PA level (24%), while no differences were noted at the AC level. Similarly, significant differences were detected for mathematics ($X^2(3, n = 8{,}131) = 781.63, p <. 05$), with the Angoff method categorising 9% more learners at the PA level, the OSS method 8% more at the AD level, and no statistical differences noted at the other PLs. Overall, the findings show that for both English and mathematics, cut scores from the Angoff and OSS methods categorise large proportions of learners at the PA and NA levels and smaller proportions at the AC and AD levels. This finding corroborates other studies that have shown that the majority of South African learners perform at very low levels (DBE, RSA, 2013; Mihai & Van Staden, 2019; Reddy, 2018).

A comparison across subjects indicates that, except at the PA level, mathematics cut scores generated from the Angoff method were significantly higher than those generated from the OSS method, and that the differences were of substantive magnitude. The fact that cut scores from the two SS methods at the PA levels were not significantly different may require further investigation. Otherwise, for English, the Angoff cut scores were significantly higher than OSS cut scores at all levels.

Previous studies that compared the Angoff and other SS methods reported cut scores that were significantly different (George, Haque & Oyebode, 2006; Jaeger, 1989; Stone, 2001). Schnabel (2018) compared cut scores from the Angoff and Item Mapping SS method, another Rasch-based method, and reported the Angoff-generated cut scores to be generally higher. It remains to be researched whether the Angoff method inherently predisposes panellists to rate student performance relatively higher, or whether there are other factors that could explain this phenomenon. On the contrary, OSS cut scores have been reported to be both robust and stable, chiefly because they derive from subject content rather than panel contests (Sondergeld et al., 2018).

At the practical level, the fact that the Angoff method generally generates higher cut scores than the OSS has critical implications, depending on whether the examination is a high- or low-stakes assessment. For a high-stakes examination like Grade 12, any SS method that generates higher cut scores will result in relatively lower pass rates, which would arguably receive mixed reception. Within the South Africa context, this could further exacerbate existing perceptions that setting high standards will impact negatively on previously disadvantaged groups, and further entrench inequality within the education system.

Question 3: Which Method would be More Appropriate for the South African Context?

To determine the appropriateness of each SS method to the South African context, we identified four key factors, each of which are further discussed.

Engagement with test content

Both the Angoff and OSS methods involved panellists interacting with test data to transform it into more meaningful information than is communicated in raw scores. In the Angoff, panellists rated test items to estimate scores of minimally competent borderline performers. Although the panellists were not directly invoking content knowledge and skills to inform their estimations, the information was used to strengthen arguments in support of their ratings for each item. In the OSS method, panellists interacted to evaluate the content that was represented by identifying essential items at each PL. Then, through an incisive "skills audit" process, the core content at each PL was interrogated to distil the specific knowledge and skills that characterise learners who, according to their total scores in the test, are deemed to function at that particular PL. For both methods, the final standard reported was a quantitative score, accompanied by a detailed description of what the test-takers knew and were able to do.

Resource requirements

The SS exercises for the Angoff and OSS methods were undertaken at different times given the availability of panel members and facilitators. The Angoff process lasted for 2 days (16 hours), requiring the costs of overnight accommodation, land and air travel expenses for some participants; as well as meals, venues, and materials to be covered. In addition, technical support was required from two researchers with specialised expertise and experience. A large amount of time was spent training panellists in the iterative item rating and discussion processes.

The OSS session lasted for one half day (4 hours); although panellists were provided with a light snack such a session would not require overnight accommodation for those travelling within the country. The training took approximately an hour i.e., 25% of a session, where one senior researcher was adequate to provide guidance and support. The majority of the time was spent on helping participants understand the concept of "essential" items. Once participants understood this concept, the actual process of rating items was completed in approximately an hour. The data entry and analysis were completed the following day by the senior researcher, who possessed the requisite expertise in the use of the IRT software. In this instance, while the participation of the panellists took a relatively short amount of time, the majority of the technical work was completed by the specialist researcher after the SS session.

Required expertise and software

For the OSS method, specialised technical expertise and experience in the use of IRT or Rasch-based analysis was required. Within the context of South Africa, this expertise is not readily available. In addition, access is also required to specialised software which is costly to obtain. However, it must be acknowledged that the use of freeware from applications such as R (Rizopoulos, 2006) could have offset some of these costs. For the Angoff method, the technical demands were substantially less, while analysis was conducted using Microsoft Excel, an application that is readily available.

Collective accountability

For the Angoff method, the entire SS processes, i.e., review of items and calculation of cut scores, was completed during the SS workshop in the presence of, and in collaboration with, panel members. Moreover, panel members were also provided an opportunity to review and comment on the final cut scores. All participants were intricately involved in the entire process and could assume full accountability for both the process and the outcome. For the OSS method, the specialist researcher had to complete the final analysis after the SS workshop, while there was no opportunity for panel members to review and discuss the final cut scores. In this instance, the responsibility and accountability for the final cut score were in the hands of the single researcher.

Conclusion

This study was prompted by observations of conceptual shortcomings in the CAPS reporting framework in South Africa. We argued that SS methods provide a more meaningful alternative. We then compared the Angoff method and the OSS method to determine which method would provide a more appropriate option for the South African context. The main finding confirms what has been observed elsewhere, namely that different methods of SS typically generated different cut scores, even when, unlike in our study, the same panels were involved (Clauser, Harik, Margolis, McManus, Mollon, Chis & Williams, 2008; Jaeger, 1989). In our study, the two SS methods differed in substantial aspects of interest in the South African context pertaining to process, outcomes and resource requirements.

In terms of process, the Angoff method focuses on predicting performance while the OSS prioritises evaluating subject content. The cut scores generated from the Angoff procedure were consistently higher than those generated from the OSS. The implications in terms of the ensuing categorisation of learners across PLs largely depend on the SS method used, a finding that was also made in other SS comparative studies in the United Kingdom and Australia (MacDaugall, 2015; Ward, Chiavaroli, Fraser, Mansfield, Starmer, Surmon, Veysey & O'Mara, 2018). However, the Angoff method tended to be more resource-intensive, a feature that cannot be ignored in a context where resources are unequally distributed. From an educational perspective, the OSS method seemed to be an ideal option, because it generates standards that are rich in curriculum information and are, therefore, more likely to add value in helping improve teaching and enhance learning in schools.

On the one hand, concerns arose when we considered that adopting the OSS method could compromise highly valued principles of access, broad participation and collective accountability in the South African education context. The OSS method required a high level of technical expertise, involved the use of specialised and costly software, and limited opportunities for collective decision-making regarding the final outcome. On the other hand, adopting the Angoff method could offset some of these concerns, because it works with easily available and less costly software that many teachers and officials are familiar with, and facilitates greater accountability for the final outcome. The use of the Angoff method incurs inordinately high costs, mainly through necessary training and attendant logistics. De Lisle (2015) also reported on the impact of a lack of capacity, financial resources and professional expertise on the SS process in Trinidad and Tobago.

What these findings point to is the need for judicious choices to be made that account for the specific context within which SS approaches are applied. More importantly, however, is the need to develop and promote a culture of data use and specifically the use of SS to improve the quality of assessment reports and their use in enhancing teaching and learning.

Our next step is to investigate the practical implications of using SS approaches for district officials, schools and teachers. In this regard, the challenge would be to collaborate with at least one education district to share experiences, provide support and help promote a culture of data use for decision-making. Capacity building in the area of SS and generating standards-based reports that users will find more meaningful and thus enhance its use to improve teaching and enhance learning will comprise the primary objectives. Moreover, these initiatives can also support current plans of the DBE to enhance the use of assessment for addressing the challenges of equity and quality in South African schools (Chetty, 2019; Mweli, 2018).

Acknowledgements

Funding for this study was obtained from the Assessment for Learning Niche Area, Tshwane University of Technology.

Authors' Contributions

AK led the conceptualisation of the paper. QM did most of the literature review and provided the tables and figures. Both authors contributed to the writing and revision of the manuscript.

Notes

i.

ii. DATES: Received: 20 November 2019; Revised: 23 July 2020; Accepted: 24 September 2020; Published: 30 November 2021.

References

Angoff WH 1971. Scales, norms, and equivalent scores. In RL Thorndike (ed). *Educational measurement*. Washington, DC: American Council on Education.

Baird JA, Isaacs T, Opposs D & Gray L (eds.) 2018. *Examination standards: How measures and meanings differ around the world*. London, England: IOE Press.

Bejar II 2008. Standard setting: What is it? Why is it important. *R&D Connections*, 7:1–6. Available at https://www.ets.org/Media/Research/pdf/RD_Connections7.pdf. Accessed 13 November 2021.

Bond TG & Fox CM 2007. *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Chetty M 2019. *The Department of Basic Education's perspective on GET assessment*. Paper presented at the Mpumalanga Department of Education School Based Assessment, Witbank, South Africa, 29–30 July.

Cizek GJ & Bunch MB 2007. *Standard setting: A guide to establishing and evaluating performance standards on tests*. London, England: Sage.

Clauser BE, Harik P, Margolis MJ, McManus IC, Mollon J, Chis L & Williams S 2008. An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education*, 22(1):1–21. https://doi.org/10.1080/08957340802558318

Cohen J 1992. A power primer. *Psychological Bulletin*, 112(1):155–159. https://doi.org/10.1037/0033-2909.112.1.155

De Lisle J 2015. Installing a system of performance standards for national assessments in the Republic of Trinidad and Tobago: Issues and challenges. *Applied Measurement in Education,* 28(4):308–329. https://doi.org/10.1080/08957347.2015.1062765

Department of Basic Education 2011. *Curriculum and Assessment Policy Statement (CAPS)*. Pretoria, South Africa: Government Printers.

Department of Basic Education, Republic of South Africa 2013. *Report on the annual national assessment of 2013: Grades 1 to 6 & 9*. Pretoria: Author. Available at https://www.education.gov.za/Portals/0/Documents/Reports/ANA%20Report%202013.pdf?ver=2014-02-07-112211-420. Accessed 13 November 2021.

Department of Basic Education, Republic of South Africa 2017. *The development of a systemic evaluation model for the basic education sector*. Available at https://peuoffice.com/wp-content/uploads/2017/11/ENDORSED-SYSTEMIC-EVALUATION-DOCUMENT.pdf. Accessed 13 November 2021.

Fritz CO, Morris PE & Richler JJ 2012. Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1):2–18. https://doi.org/10.1037/a0024338

George S, Haque MS & Oyebode F 2006. Standard setting: Comparison of two methods. *BMC Medical Education*, 6:46. https://doi.org/10.1186/1472-

6920-6-46

Griffin P & Nix P 1990. *Assessment and reporting: A new approach*. Sydney, Australia: Harcourt Brace Jovanovich.

Haertel EH 2005. Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice*, 21(1):16–22. https://doi.org/10.1111/j.1745-3992.2002.tb00081.x

Hambleton RK & Pitoniak MJ 2006. Setting performance standards. In RL Brennan (ed). *Educational measurement* (4th ed). Westport, CT: Praeger.

Jaeger RM 1989. Certification of student competence. In RL Linn (ed). *Educational measurement* (3rd ed). New York, NY: Macmillan.

Kanjee A, Claassen N, Makgamatha M & Diedricks G 2004. *Grade nine learner achievement monitoring programme: Technical report* (Unpublished report). Pretoria, South Africa: Human Sciences Research Council.

Kanjee A & Moloi Q 2016. A standards-based approach to reporting assessment results in South Africa. *Perspectives in Education*, 34(4):29–51. https://doi.org/10.18820/2519593X/pie.v34i4.3

Linacre JM 2014. *Winsteps® Rasch measurement computer program*. Beaverton, OR: Winsteps.com.

MacDaugall M 2015. Variation in assessment and standard setting practices across UK undergraduate medicine and the need for a benchmark. *International Journal of Medical Education*, 6:125–135. https://doi.org/10.5116/ijme.560e.c964

Mihai M & Van Staden S 2019. Ondervindinge, uitdagings en suksesse: Vroeëleesbegrippraktyke in hulpbronbeperkte omgewings met kinders uit linguisties-diverse agtergronde [Experiences, challenges and successes: Early-reading comprehension practices in resource-constrained settings with children from linguistically diverse backgrounds]. *Tydskrif vir Geesteswetenskappe*, 59(3):436–450. https://doi.org/10.17159/2224-7912/2019/v59n3a8

Moloi M & Kanjee A 2018. Beyond test scores: A framework for reporting mathematics assessment results to enhance teaching and learning. *Pythagoras*, 39(1):a393. https://doi.org/10.4102/pythagoras.v39i1.393

Monahan PO & Ankenmann RD 2005. Effect of unequal variances in proficiency distributions on Type-I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement*, 42(2):101–131. https://doi.org/10.1111/j.1745-3984.2005.00006

Mweli M 2018. *Improving assessment practices*. Paper presented at the Director General's meeting with the Kwazulu-Natal Provincial Education Department, Durban, South Africa, 20 April.

Nässtrőm G & Nyström P 2008. A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research & Evaluation*, 13(9):1–12. https://doi.org/10.7275/bhb9-8t88

Pitoniak MJ & Yeld N 2013. Standard setting lessons learnt in the South African context: Implications for international implementation. *International Journal of Testing*, 13(1):19–31. https://doi.org/10.1080/15305058.2012.741085

Reddy V 2018. *TIMSS in South Africa: Making global research locally meaningful*. Available at http://www.hsrc.ac.za/en/review/hsrc-review-april-june-2018/timss-in-sa. Accessed 28 January 2019.

Rizopoulos D 2006. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25. https://doi.org/10.18637/jss.v017.i05

Scherman V, Zimmerman L, Howie SJ & Bosker R 2014. Setting standards and primary school teachers' experiences of the process. *Perspectives in Education*, 32(1):92–104.

Schnabel SD 2018. A comparison of the Angoff and item mapping standard setting methods for a certification examination. PhD thesis. Chicago, IL: University of Illinois at Chicago. Available at https://indigo.uic.edu/articles/thesis/A_Comparison_of_the_Angoff_and_Item_Mapping_Standard_Setting_Methods_for_a_Certification_Examination/10941725. Accessed 13 November 2021.

Sondergeld TA, Stone GE & Kruse LM 2018. Objective standard setting in educational assessment and decision making. *Educational Policy*, 34(5):735–759. https://doi.org/10.1177/0895904818802115

Stone GE 2001. Objective standard setting (or truth in advertising). *Journal of Applied Measurement*, 2(2):187–201.

Tiratira NL 2009. Cutoff scores: The basic Angoff method and the Item Response Theory method. *The International Journal of Educational and Psychological Assessment*, 1(1):27–35.

Ward H, Chiavaroli N, Fraser J, Mansfield K, Starmer D, Surmon L, Veysey M & O'Mara D 2018. Standard setting in Australian medical schools. *BMC Medical Education*, 18:80. https://doi.org/10.1186/s12909-018-1190-6

Zieky M & Perie M 2006. *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service (ETS). Available at https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf. Accessed 13 November 2021.