An Exploratory Study on the use of Two Standard Setting Methods in the Validation of Mother Tongue Language Descriptors – The Singapore Experience

Khee Shoon, TEO (teo_khee_shoon@seab.gov.sg)
Yi An, SOH (Ms) (soh_yi_an@seab.gov.sg)
Cheow Cher, WONG (wong_cheow_cher@seab.gov.sg)
Lay Keng, CHUA (Ms) (chua_lay_keng@seab.gov.sg)

Singapore Examinations and Assessment Board, Singapore

Abstract

To nurture students to be proficient users of the Mother Tongue languages in Singapore, a proficiency-oriented approach has been recommended. Central to the proficiency-oriented approach is a set of language proficiency descriptors that explicitly spell out the six core language skills (Reading, Listening, Speaking, Spoken Interaction, Writing and Written Interaction) and the levels of attainment at key stages of learning. Such proficiency descriptors could help guide teachers in their teaching and inform students about their learning progress.

An exploratory study was conducted to validate the proficiency descriptors. In this study, the 'Bookmark method' and 'Body of Work' standard setting methods were used, and feedback into the validation cycle with further triangulation through teacher surveys and interviews regarding the proficiency descriptors. This paper shares some of the experiences and learning from the application of these standard setting methods in the process of validating the proficiency descriptors; in particular, the importance of building consensus and clarifying standards amongst the standard setting panelists.

Keywords: Validation, Proficiency Descriptors, Body of Work Method, Bookmark Method

Introduction

The Singapore Mother Tongue Language (MTL) curriculum aims to nurture students into proficient language users who can use their MTL to communicate in a confident, effective and meaningful manner in everyday life. As students from different home backgrounds enter elementary schools with different starting points, the curriculum needs to cater to their varying learning needs. Hence, clear descriptions of what students are able to achieve at different proficiency levels have to be spelt out explicitly in the MTL curriculum. The descriptors serve as a milestone checkpoint to guide students and teachers on what is expected at the key stages of learning and to motivate the progress of students from one proficiency level to the next. Hence, a proficiency-oriented curriculum has been recommended for the MTL curriculum.

The proficiency-oriented approach comprises the actions performed by learners to develop an appropriate range of language competences (i.e. language knowledge, language skills and language strategies). Central to the proficiency-oriented approach is a set of proficiency descriptors that explicitly spell out the six core language skills[1] and levels of

[1] Listening, Speaking, Reading, Writing, Spoken Interaction and Written Interaction

attainment at key stages of learning.[2] The overall proficiency descriptors are unpacked into various 'Can Do' statements, which describe what language users can typically do with the language at different stages of learning, using various language skills in a range of contexts. The proficiency descriptors and 'Can Do' statements for Level 4 Reading are shown in **Annex A**. To validate the proficiency descriptors and to ascertain if the proficiency descriptors have appropriately spelt out the levels of attainments students should achieve at various key stages of learning, a process is established to validate through empirical data.

Validation Approach

A typical validation approach consists of three phases, namely a) Task development; b) Field testing; and c) Standard Setting. The results are then interpreted and recommendations are made for the necessary follow-up actions, such as reviewing the 'Can Do' statements, the attainment level of specific group of students and the tasks used for validation (see **Annex B**).

Task Development

To validate a particular proficiency descriptor, the 'Can Do' statements were first translated into an assembly of language tasks designed to elicit students' responses so that they could demonstrate their language competence at that specified proficiency level. Based on the proficiency descriptors and corresponding 'Can Do' statements as well as the task parameters (e.g. range of vocabulary, grammar structure), a set of validation tasks was developed. For this exploratory study, the validation tasks for Levels 3 to 5 of the six core language skills (including oral and written interaction skills) were developed.

Field Testing

The validation tasks were administered to representative samples of Grade 6 students and Grade 7 (as proxy to Grade 6) in late Nov 2011 and early Feb 2012 respectively. About 1,600 Grade 6 pupils from 44 primary schools and 1,700 Grade 7 pupils from 42 secondary schools sat for the various tests validating 6 core skills at 3 proficiency levels. The scoring process involved rigorous pre-field test preparations and standardisation to ensure accuracy and reliability in marking.

Standard setting

Students' artefacts from the field tests and validation tasks developed were examined by expert panels to determine if students were able to achieve the demands spelt out in the language tasks and attain the language proficiency at a specified level. Professional judgment from the expert panels was collected through standard setting exercises. Bookmark method was adopted for the validation of receptive skills (Listening and Reading Skills) whilst Body of Work method was adopted for the validation of productive skills (Speaking, Spoken Interaction, Writing and Written Interaction). The choice of these two methods was determined in view of the differences in the nature of the tasks, as supported by literature scans that the use of Bookmark Method is for multiple choice questions and short answer type responses whilst that of Body of Work Method is for holistic responses such as essays and oral responses.

[2] Grade 2, Grade 4, Grade 6, Grade 8, Grade 10 and Grade 11/Grade 12

Focus group discussions with teachers

In addition to collecting data from standard setting sessions, qualitative inputs were gathered from practitioners. Focus group discussions involving 64 teachers were conducted to elicit feedback and teachers' perceptions on the proficiency descriptors exemplars pegged at the respective levels. These teachers were from different school types and levels. They were asked to determine the proficiency levels of reading and listening tasks and explain their choices. Data such as teachers' reasons and views on why they pegged exemplars at the respective levels were collected to provide some reference to the appropriateness of the levels. The findings from the FGDs were then subsequently triangulated with the outcomes from the standard setting sessions to further validate the proficiency descriptors.

Key Considerations for Standard Setting

Several key considerations were taken into account in the standard setting sessions.

Selection of standard setting panelists

The particular group of persons selected as standard setting panelists and the training these panelists receive were important in the standard setting process as the outcomes could affect the eventual standards recommended.

For this study, the standard setting panel was made up of curriculum planning specialists, assessment specialists and master teachers who were leading practitioners in their subject discipline. Panelists were selected based on their content knowledge, expertise and experience in the existing curriculum and the MTL proficiency descriptors and 'Can Do' statements. They have kept abreast with the current skill sets and knowledge students had and a good understanding and expectation of the students' language competencies.

Training of standard setting panelists

After the careful selection of the panelists, training was provided to ensure the panelists had a clear understanding of their roles during the standard setting process.

Overview of the Bookmark method

The bookmark method was one of the standard setting methods employed during the standard setting session for the skills of *Listening* and *Reading*. It was designed to yield scores on the basis of expert panelists' reviews of test questions (Lewis, Mitzel, and Green, 1996, Mitzel, Lewis, Patz, Green, 2001). The method was named "Bookmark" because panelists had to place markers within the *Ordered Item Booklet* (OIB), which consisted of a set of items arranged from easiest to most difficult.

Before the start of the standard setting session, intensive preparation was required to prepare the OIB which displayed items in order of difficulty from the easiest to most difficult. These items' difficulty values were calibrated according to Item Response Theory (IRT) based on the empirical data collected from the field tests. The IRT estimates of difficulty values and students' ability estimates were placed on the same scale.

One practicality of the OIB is that it could be used to display items which were in multiple-choice format and/or constructed response format, and suitable for test like Reading which consists of a mixture of formats

Response Probability Value and Borderline student

It is important that panelists understood the concepts of "Response Probability" values and "Borderline" students before the start of the standard setting session. A Response Probability (RP) value of two-thirds or 0.67[3] was chosen before the standard setting procedure starts. This RP value referred to the probability of the "borderline" student having 67% likelihood or 2/3 chance of answering the item correctly. The "borderline" student referred to an abstractly defined student deemed to possess the minimum competencies and abilities required at the specified proficiency level.

Implementation of the Bookmark method

Before the standard setting session, panelists were given time to familiarize themselves with the proficiency descriptors and the 'Can Do' statements. They also received the item ordered booklet and recording sheet to record their bookmarks.

During round one of the standard setting session, panelists started with the easiest item in the OIB to determine, for each item, whether the borderline students were able to answer at least 2 out of 3 similar items correctly. If the answer was "yes", they moved on to the next item in the *OIB* until they reached the item where they considered the answer to be "no". A bookmark was then placed at the first item where panelists, in their opinion, felt the borderline student had a less than two-third probability of answering the item correctly. Panelists would first place their individual bookmark placement during this round.

Next, they would share their bookmark placements within the assigned group and the reasons for the bookmark placements. Panelists may choose to or choose not to change their individual placement of the bookmarks. During round 2 of the discussion, everyone would discuss and share their bookmark placements together. The cut-score computations using IRT would then be carried out based on the bookmark placements.

Learning gained

"Probability Value" and "Borderline" Concepts

It was observed that in the beginning during the standard setting sessions, panelists experienced some difficulty in applying the concept of probability value. Having to imagine borderline students able to answer the question correctly with a probability of 0.67 was an abstract task and posed a challenge for some of them, especially for constructed response items, where panelists imagined borderline students able to obtain *at least* x marks for a constructed response item with a probability of 0.67. Hence, to reduce the cognitive load of panelists, facilitators encouraged panelists to imagine students being able to answer at least 2 out of 3 items (all measuring at the *same particular difficulty level)* correctly. Alternatively, panelists may interpret the RP value as the probability of two-thirds of the borderline students being able to answer the particular item correctly. Panelists appeared more comfortable applying the probability response concept for the subsequent standard setting sessions after this concept was clarified and elaborated.

[3] Response probabilities other than 0.67, such as 0.50 and 0.80 have been used, but according to literature review, 0.67 is the most commonly used response probability for Bookmark method studies.

It was also observed that in the beginning, panelists had some difficulties internalizing the concept of "borderline" student. Some misinterpreted them as the average performing students. In fact, this borderline student referred to one who had minimal competencies in achieving the 'Can Do' statements of Reading, for example. It was not unexpected that every panellist had a different notion of what "borderline student" meant to them as each one had varied teaching experience in different schools and hence might hold different expectations of borderline students in terms of their skills and competencies. To address this issue, facilitators first explained it was difficult for everyone to hold a common yardstick of what "borderline" student meant but encouraged all panelists to discuss and articulate what "borderline" meant to them and the skills and competencies "borderline" student had with reference to the 'Can Do' statements. Through the discussion and clarification, a more common understanding was reached on what "borderline student" meant. This could help ensure that the standard setting sessions were conducted in a more consistent manner.

Ordered Item Booklet

During some of the standard setting sessions, there were instances where panelists did not agree with the order of the items in the booklet. It must be made clear to panelists that the order of the items was based on statistical calculations of difficulties, not the judged difficulties of the items. Also, panelists were reminded that in certain parts of the booklet, adjacent questions may be very close to each other in difficulty while in some parts, there may be questions which difficulty varied greatly from each other. Panelists should not expect the difficulties of the items to be equally distributed in the booklet.

It was not surprising that panelists' judgments of their bookmark placements differed. Facilitators should encourage panelists to discuss the reasons why they placed bookmarks at a particular location and list out the core knowledge and skills expected for each task. Discussion was critical here to ensure panelists reached a common understanding of the cognitive requirements for the task.

Overview of the Body of Work method

The Body of Work method was another standard setting method employed during the standard setting session for the productive skills of *Writing, Written Interaction, Speaking* and *Spoken Interaction*. This method was designed for assessments which comprised of constructed responses that yield observable students' products such as essays and oral responses (Kahl, Crockett, DePascale, & Rindfleisch, 1994, 1995). It was a holistic standard-setting method that required expert panelists to evaluate whole sets of examinee work to help to pinpoint a cut score in a systematic manner (Kingston, Kahl, Sweeney, & Bay, 2001). The next few paragraphs below detailed the procedure for the Body of Work method carried out for the productive skills. The term "scripts" below would refer to the students' written responses to the *Writing* and *Written Interaction* tasks, and to the students' oral responses to the *Speaking* and *Spoken Interaction* tasks.

Implementation of the Body of Work method

Prior to the standard setting session, a folder called a *Range-Finding Folder* was prepared. It consisted of about 8 to 10 students' scripts. These scripts were sampled from the entire range of awarded scores at regular intervals. This was to provide the panelists with the entire range of performance and ability exhibited by the students in response to the given

tasks. Before the actual standard setting session began, panelists were asked to familiarize themselves with the 'Can Do' statements and the items which assessed the 'Can Do' statements; as well as the marking rubrics that were used to score the scripts of the students. By examining the 'Can Do' statements and the descriptions in the marking rubrics, the panelists could form an initial estimate of where the cut-score is likely to be.

During the standard setting session, panelists individually examined each sampled script in the *Range-Finding Folder*, without knowing the score that has been awarded to each script. For each sampled script, they individually judged whether the student was deemed to possess the minimum competencies and abilities required at the specified proficiency level. After the first round of individual evaluation, panelists shared with the group their individual verdicts of each script in the *Range-Finding Folder*. The scores awarded to each script were then revealed to the panelists. After another round of discussion based on the verdicts of each script and its score, the panelists decided on a narrower range of scores where the cut-score was likely to be, for example, between 16 to 20 marks, based on their verdicts of the 16-mark and 20-mark scripts. This round, during which the narrower range of scores has been identified, is known as the *Range-Finding Round*.

After identifying the narrower range of scores, a folder known as a *Pin-pointing Folder*, consisting of 3 scripts sampled from each score-point in the identified range, was given to the panelists. The panelists would then carry out a similar procedure as described in the *Range-Finding Round*. After final discussion of where the cut-score was deemed to be, the panelists' individual judgments were collated. The cut-score would be established if there was consensus among the panelists; otherwise, the cut-score computations would then be carried out using logistic regression, based on the judgments collated in the final round.

Learning points

Several learning insights were gained when the Body of Work method was implemented. While preparing the materials for the standard setting session for Speaking and Spoken Interaction, for example, careful selection of audio clips for range-finding round and pin-pointing round had to be made. Once the range of marks for the range-finding round was determined, facilitators had to start the intensive preparation of listening and selecting audio clips which were of good quality, i.e. audible and clear to the listeners. This was because some students' voices might be so soft that they might not be clearly audible from the recordings. As for preparing the materials for Writing and Written Interaction, facilitators had to ensure the sample scripts selected had at least legible handwriting for the panelists to decipher what was written, and had "face validity" for the marks awarded to that sample script. The effort and time spent during this preparation stage would be worthwhile as it could minimize frustration and distraction among panelists during the actual conduct of the standard setting session.

It was also observed that initially in the beginning of the standard setting session, while judging if the selected samples were able to meet the 'Can Do' statements in the proficiency descriptors, some panelists referred to the marking rubrics and had the tendency to focus too much on one particular aspect of the rubrics, for example, Content for Writing, in which the components assessed for Writing were Content and Language. Facilitators would need to highlight to panelists that the judgment of students' scripts were holistic in nature and should not be judged based only on specific criteria. Also, panelists had to take into account the 'Can Do' statement of the particular descriptors when judging the quality of the scripts.

After a few rounds of standard setting sessions, panelists became more familiar with the procedures and gradually gained a better understanding on the requirements of the Body of Work procedure.

Some of the learning gained from conducting the Bookmark method could also be applied to the Body of Work method. For example, participants may have disagreements on whether a particular script was able to meet the 'Can Do' statement. Facilitators could encourage discussion among panelists to share the reasons why they agree/disagree and list out the knowledge and skills expected for that question. Through this discussion, a more common understanding of the cognitive requirements for the question could be reached among participants.

Conclusion

This paper has discussed the validation approach and the learning points from the conduct of standard setting exercises. The key findings of this study provided useful base-line data to guide the team in validating and refining the proficiency descriptor framework. The findings also served as useful input for developing the new instructional packages, which would provide greater clarity on proficiency expected for the different language skill sets.

As for the validation of the proficiency descriptors, there is no perfect standard setting method. To ensure that the standard setting process is appropriately carried out, there are many important considerations to bear in mind, such as the selection of panelists, preparation, training, providing panelists with relevant data to make good judgements and ensuring consistency in the conduct of standard setting sessions.

References

Kahl, S.R., Crockett, T.J., DePascale, C.A., & Rindfleisch, S.L. (1994, June). *Using actual student work to determine cutscores for proficiency levels: New methods for new tests.* Paper presented at the National Conference on Large-Scale Assessment, Albuquerque, NM.

Kahl, S.R., Crockett, T.J., DePascale, C.A., & Rindfleisch, S.L. (1995, June). *Setting standards for performance levels using the student-based constructed-response method.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kingston, N.M., Kahl, S.R., Sweeney, K.P., Bay, L. (2001). Setting Performance standards Using the Body of Work Method. In G. J. Cizek (Ed.), *Setting Performance Standards, Concepts, Methods, and Perspectives*. Mahwah, N.J.

Lewis D.M., Mitzel, H. C., Green, D. R. (1996). Standard Setting: A Bookmark Approach. In D. R. Green (Chair), *IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring*. Symposium presented at the 1996 Council of Chief State School Officers 1996 National Conference on Large Scale Assessment, Phoenix, AZ.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

Annex A

Proficiency descriptors for Reading

Level/ MTL	Overall Language Proficiency Descriptors for Reading	**Can-Do Statements** Understand a variety of written texts for different purposes		
		1.Narrate, describe (Narrative text)	**2.Inform, explain (Informative text)**	**3.Express views, convince (Persuasive text)**
Level 4 Reading	I can understand written texts on topics related to self, family, school and community. The texts employ common organizational structures, use basic vocabulary and common sentence structures.	I can understand written texts and how most details come together to form the theme. I can evaluate actions of characters.	I can understand information and details in written texts.	I can understand the author's opinion and reasons in written texts.

Annex B

Figure 1: Validation Approach for MTL Proficiency Descriptors