Chapter 5
Standard Setting in PISA and TIMSS and How These Procedures Can Be Used Nationally

Rolf Vegar Olsen and Trude Nilsen

Abstract In this chapter, we compare and discuss similarities and differences in the way the two large-scale international studies, PISA and TIMSS, formulate and set descriptions of standards. Although the studies use similar methods, different decisions have been made regarding the nature and properties of the final descriptions of student achievement. In addition to this overview, we treat PISA and TIMSS as case studies in order to illustrate an under-researched area in standard setting: the nature of and empirical basis for the development of performance level descriptors (PLDs). We conclude by discussing how these procedures may be relevant for formulating useful standards in tests and assessments in the Norwegian context.

Keywords Performance level descriptors • Standard setting • Large-scale assessment • PISA • TIMSS

5.1 Introduction

One of the more powerful ways to report the PISA 2000 scores was to use performance level descriptions (PLDs). The state of shock communicated by policymakers in several countries following the presentation of the PISA results may have been caused in part by the power of these descriptions. For instance, policymakers were warned that "…[E]ducation systems with large proportions of students performing below, or even at, Level 1 should be concerned that significant numbers of their students may not be acquiring the literacy knowledge and skills to benefit sufficiently from their educational opportunities" (OECD 2001, p. 48).

R.V. Olsen (✉)
Center for Educational Measurement, University of Oslo, Oslo, Norway
e-mail: r.v.olsen@cemo.uio.no

T. Nilsen
Department of Teacher Education and School Research, University of Oslo, Oslo, Norway
e-mail: trude.nilsen@ils.uio.no

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_5

In short, the scales were partitioned into a finite number of intervals, and information about students' relative success on test items was used to develop verbal descriptions characterizing students' performance as they progressed on the scale. PISA was not the first large-scale assessment to develop and implement these types of descriptions—similar procedures were developed and applied successfully in both the National Assessment for Educational Progress (NAEP) (Beaton and Zwick 1992) and TIMSS (Kelly 1999).

With some exceptions concerning national assessments[1], standard setting rarely occurs in Norway. However, teachers and exam judges are given the task of grading students, and at least at a superficial level, the end product of grading resembles the end product of standard setting procedures, because they both consist of a limited number of levels or cut scores that are intended to represent a coarse measure of the student achievement. In other words, some rules or procedures that are applied result in grades; however, there is very little understanding of what the grades actually represent or of teachers' reasoning when making grading decisions.

In this chapter, we discuss the similarities and differences in the way the two large-scale international studies, PISA and TIMSS, formulate and set their descriptions of standards. In doing so, we also briefly relate these procedures to the wider literature on standard setting (e.g., Cizek 2012; Cizek and Bunch 2007; Smith and Stone 2009). Previous studies emphasized various aspects of how to use expert judgments to identify substantially meaningful cut scores along a scale representing the measurement of achievement in a specified domain. Here, we investigate how decisions are operationalized in large-scale international studies and extend the discussion to a less researched area: the nature of and basis for the development of the descriptions of student achievement along the scale. These are potentially powerful tools for communicating the results of the studies. In concluding, we discuss how these procedures may be relevant to conceiving and operationalizing useful standards in assessments in the Norwegian context.

5.2 PISA and TIMSS: Differences and Similarities

Before describing how PISA and TIMSS produce their descriptions of students' proficiency at different points along the scales, it is necessary to give a short account of how the two studies differ. To some extent, the nature of the final descriptions of students' proficiencies could be regarded as reflecting the somewhat different perspectives and aims guiding the two studies. We use science as the example domain in this chapter; hence, we refer to some specific aspects of how the two studies have

[1] The national assessment is comprised of compulsory reading, English and numeracy tests conducted at the start of the school year as students enter upper primary school (5th grade) and lower secondary school (8th grade). They are low-stakes assessments meant to be used formatively for students, but they are also used for accountability purposes for schools.

defined and operationalized this domain. Similar statements could, however, be made for mathematics.

Both TIMSS and PISA include measures of students' competencies in science and mathematics. In addition, PISA includes a measure of reading competency and one so-called innovative domain varying from cycle to cycle (e.g., collaborative problem solving). The major important difference between the assessments is that while the TIMSS framework and design is firmly based on a model of school curriculum (Mullis et al. 2009), PISA is based on a more future-oriented perspective that seeks to identify knowledge and competencies needed for further studies, careers, and citizenship in general, emphasizing what could be termed a *systems perspective* (OECD 2006). Consequently, TIMSS samples intact classes in order to study instructional processes, whereas PISA samples students across classes within schools. Also, TIMSS includes grades in the middle of primary school (4th grade) and the beginning of lower secondary school (8th grade), whereas PISA samples an age cohort toward the end of compulsory schooling (students turning 15 in a specific year).

The tests are constructed quite differently in the two assessments. PISA uses clusters of items with a common stimulus material, often in the form of an extended text, while TIMSS mainly contains stand-alone items, including "pure or context-free" items. While TIMSS places equal emphasis on science and mathematics in each survey, PISA has a system in which one of the three core domains is allocated more time every third cycle. One consequence of these differences in how the tests are constructed is that TIMSS includes a far greater number of total items in each domain. For instance, when science was the major domain in PISA 2006, a total of 109 items covered the domain, while TIMSS always has more than 200 items in each of the two domains.

There are other similarities and differences between the two assessments, but the aforementioned are the most relevant differences in terms of factors that directly or indirectly affect the standards they have developed (for a more detailed comparison of the two assessments, see Olsen 2005). In oversimplified terms, while PISA has a more future-oriented goal closely related to monitoring the sustainability and development of society, TIMSS aspires to study the educational effectiveness of factors proximal to what happens within classrooms.

5.3 Standard Setting Procedures

In education, the term *standard* refers to a range of different phenomena. First, standards are often used to refer to formulations of expectations. In official curriculum documents, the intended aims of the education system are described through content or competency standards. In some countries or jurisdictions, schools have to meet expectations of average performances to be achieved, and at the system level, expectations of future performance may also be defined in relation to international surveys. These expectations are often referred to as *standards* or *benchmarks*. Another use of the term *standard* refers to agreed-upon quality criteria for certain objects, such as in standards for teaching, standards for assessments, etc.

Here, we refer to a family of meanings that are related to both standards as expectations and standards as quality descriptions. Both types of standards are related to measuring performance, proficiency, or achievement within some domain of relevance for education (such as science). For our purposes, standard setting may in its widest sense be defined as "…the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (Cizek 1993, p. 100). In other words, standard setting refers to procedures that are implemented in order to identify points or intervals along a scale designed to measure student achievement within a specified domain. In what follows, the associated verbal descriptions of these points or discrete levels along the scale are regarded as parts of the standard setting procedure.

In the literature, these are often referred to as *achievement* or *performance level descriptors* or *PLDs* (Egan et al. 2012; Perie 2008). Over the last decades, standard setting has emerged as a response to several types of questions or purposes. First, standard-setting procedures have been initiated in order to provide a more rational and judicial basis for pass/fail decisions. This could, for instance, be for certification purposes aiming to ascertain that persons entering into a profession meet a standard regarded as appropriate. In this case, the procedure would involve identifying a specific cut score on an assessment. Second, particularly in the US context, standard setting serves to promote and develop criterion-based assessments, as opposed to a simple reference to a norm or a distribution. Numbers in the form of percentage correct, percentile ranks, etc. alone do not communicate what students know or are able to do—they simply state that a student is relatively more or less able as compared to a distribution of items or students. Third, in systems with several exam providers, such as the UK, regulatory processes have been installed to ensure that the exams are comparable across the different providers; this process is also referred to as *standard setting*.

Many education systems seek to ensure that standards are maintained over time (i.e., that the numbers used to report student performance in one year have a rational basis for comparison with the apparently similar numbers used in the past and in the future). Hence, standard setting is intimately related to purposes of linking and equating scores.

Standard setting procedures usually rely on judgments by panels of content or subject matter experts (e.g., teachers). These experts are tasked with deciding where along the scale they find it meaningful (based on theory and/or tacit expert knowledge) to create a cut score. The great number of specific procedures used to organize the work of such panels may be grouped into two distinct approaches*, item-* or *person-centered*, depending on whether the procedure primarily involves judging items or test takers (for details see, for instance, Cizek and Bunch 2007; Zieky et al. 2008).

Although most of these methods were originally developed within the framework of classical test theory, they are now increasingly implemented using item response theory (IRT). In particular, for several of the item-centered methods, the advantage of using IRT is that students' proficiency is placed on the same scale as the difficulty estimates of the items, often referred to as *item maps* or *Wright maps* (see Fig. 5.1). This enables development of verbal and probabilistic descriptions of

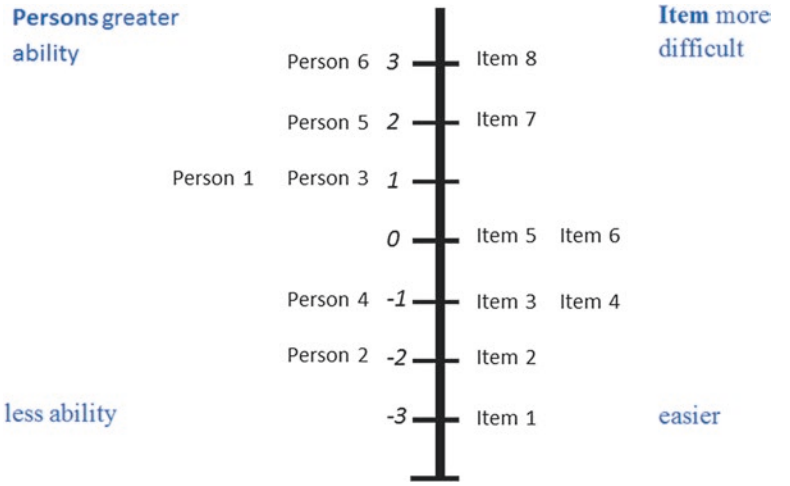

Fig. 5.1 A generic example of a person–item map

the proficiencies demonstrated by a typical student at different points or within different intervals along the scale. These methods are therefore often referred to as *item mapping* (Huynh 2009).

Figure 5.1 contains a generic and simple example of an item map. As stated above, with the help of IRT, the difficulty of the items and the ability of the students are placed along the same continuous underlying scale. The scale ranges from easier items and students with less ability at the bottom to more difficult items and students with more ability at the top of the scale. A default option in most IRT applications is to locate the items on the point of the scale where students have a 50/50 chance of succeeding. This default option is also referred to as *RP50*[2]. For example, in Fig. 5.1, Person 4 has a 50% probability of providing a correct response to Items 3 and 4 and an even higher probability of success on Items 1 and 2. However, not everyone would agree that a 50% chance of responding correctly to an item represents mastery. A somewhat stricter criterion of at least an 80% chance (RP80) could be perceived as more useful in some contexts. This adjustment is easily accommodated and would simply result in items being shifted upward in the person–item map.

5.4 Standard Setting Procedures in TIMSS and PISA

The methods used in PISA and TIMSS are based on the interpretation of person–item maps. To a large extent, the procedures applied in both studies, particularly in TIMSS, are based on those first implemented as part of the National Assessment of

[2] *RP* refers to *response probability*.

Educational Progress (NAEP) called *scale anchoring* (Beaton and Allen 1992). Both PISA and TIMSS use the following procedures:

1. Expert groups write frameworks explicating to some degree the construct being measured, including a generic hypothesized notion of the characteristics of performance from low to high on the scale.
2. Items are developed and implemented according to the specifications in the framework.
3. Item writers and expert groups develop item descriptors (IDs), which may include coding the items according to the categories used in the framework and open-ended statements with specific descriptions of the knowledge and skills involved in solving the item
4. Data are analyzed, parameters for students and items are extracted (using IRT), and graphics and tables with data (as in Fig. 5.1) are produced.
5. A (pragmatic and empirically based) decision is made about the number and location of cut scores to be used.
6. Items are identified as markers of the performance levels to be reported.
7. Performance level descriptors (PLDs) are developed based on detailed descriptions of the clusters of items identified (see Step 3) and the general description of the construct included in the framework (see Step 1)

The major difference between the procedures used in PISA and TIMSS as compared to the standard setting procedures described in the literature is that identifying cut scores is not based on a process involving a panel of judges. Instead, the practice is rooted in the premise that it is not possible or meaningful to derive substantial qualitative descriptions of thresholds along the scale from explicit or implicit theory alone. The scales are continuous and unimodal, suggesting that any cut score is equally meaningful in a qualitative sense. Decisions about the number and location of cut scores are therefore solely based on a combination of pragmatic criteria regarding usefulness for communication and empirical criteria. However, expert judgments are still vital to the process, particularly in Steps 1, 2, 3, and 7. Although PISA and TIMSS are very similar in their approach to standard setting, there are important differences between how their cut scores are developed and communicated.

Figure 5.2 provides a more detailed description of the nature of the standard setting in TIMSS and PISA. The details of the procedures are described in the technical reports (see, for instance, the latest versions: Mullis 2012 ; OECD 2014). The figure illustrates that PISA identifies more cut scores than TIMSS: six[3] and four, respectively. Another major difference is that PISA uses the cut scores to define intervals of proficiencies, while TIMSS defines what is referred to as *anchors* along the scale; these anchors are interpreted as fuzzy points. Another striking difference is that TIMSS has placed the cut scores along some preselected and well-rounded

[3] This number is typical, but five and seven cut scores have also been used in PISA.

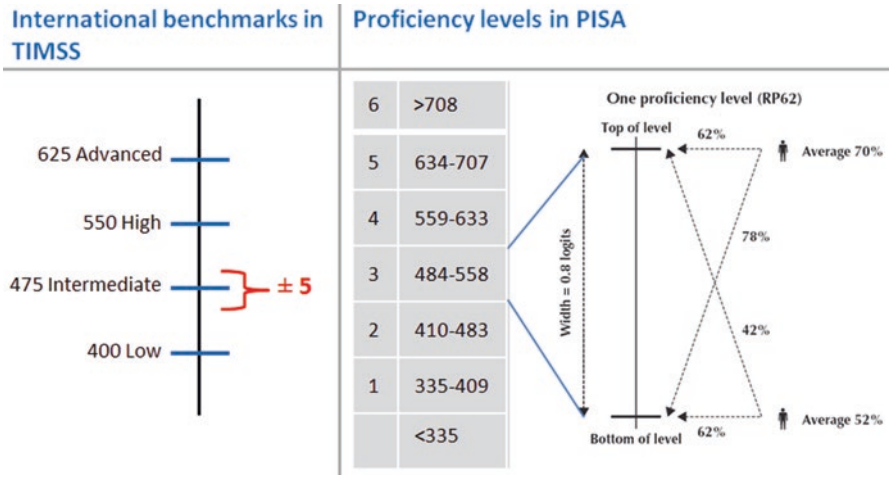

Fig. 5.2 Principles for deciding cut scores in TIMSS and PISA. The numbers for PISA refer to the science scale. Note: The right-hand figure is copied from OECD (2014, p. 293)

values on the scales, while PISA has applied another criterion for placing the cut scores, resulting in somewhat irregular values. In addition, the exact location of the cut scores is generally not equal across the domains in PISA, while TIMSS operates with the same cut scores across mathematics and science. It is interesting to note that the distance between two adjacent cut scores are rather similar in the two studies, representing about 75 points (corresponding to 75% of one standard deviation unit in the internationally pooled sample).

5.4.1 Defining the International Benchmarks in TIMSS

Standard setting is called scale anchoring in TIMSS, referring back to procedures first developed for the NAEP (Beaton and Allen 1992) and implemented for the first time in TIMSS 1999 (Gregory and Mullis 2000; Kelly 1999). Initially, these anchors (or international benchmarks) were placed on a percentile scale. However, for the 2003 assessment, the test centre realized that in order to report trends, they needed to reference defined points on the underlying scale (Gonzalez et al. 2004). The values in Fig. 5.2 have been in use ever since.

The anchoring process in TIMSS begins by identifying students who score within five scale-score points of each cut score. For these students, the percentages correct are computed for all items. Several criteria are then used to identify item anchoring at the different benchmarks. First, for a multiple-choice item to anchor at a specific benchmark, at least 65% of the students in the benchmark interval must answer it correctly. Additionally, less than 50% of the students belonging to the next

lower benchmark must respond correctly[4]. For open-ended response items, the criterion is to place the item in the lowest of the benchmarks with at least 50% correct responses[5].

5.4.2 *Defining Proficiency Levels in PISA*

The procedure used in PISA identifies a set of equally spaced intervals along the scale. The starting point for defining these levels is the idea that students at a particular level will be more likely to solve tasks at that level than to fail them. Students are therefore assigned to the highest level in which they are expected to correctly answer the majority of the assessment items. Then, a pragmatic choice is made for the width of the equally spaced intervals[6]. The last procedural step in setting up the proficiency levels in PISA is to place the lower end of Level 1 at the lowest score point possible given the requirements above. In practice, using a response probability of 62% (RP62) produces intervals with these properties. As illustrated in Fig. 5.2, by using RP62, a student at the lower end of any proficiency level is expected to give correct responses to more than 50% of the items belonging to this interval. A student at the very top of a level is expected to respond correctly to approximately 70% of the same item set.

Specific arguments about the width of the intervals applied in PISA and the distance between adjacent benchmarks in TIMSS are, to our knowledge, not explicitly documented. However, it may be reasonably assumed that these choices are affected by what is perceived to be a useful number of categories for reporting combined with the limitations given by the total number of items. The latter is important to consider because the end products of the standard-setting process in TIMSS and PISA are not a set of cut scores. Having identified these, the next step is the development of verbal descriptions of what students know and are able to do at different levels of the construct. Hence, a fair number of items are needed at each benchmark or within the proficiency levels in order to develop meaningful descriptions (Step 7 in the list above).

5.5 From Items to PLDs

As identified in the list of steps involved in the standard-setting procedures in PISA and TIMSS, items are described by IDs that reflect both generic categories used to define the construct in the framework and the specific content and cognitive demand of each item. This is the raw material used to generate PLDs. In the following

[4] This discrimination criterion for the low international benchmark could not be applied for obvious reasons.

[5] In addition, a less strict criterion was used to identify items labeled “almost anchored.”

[6] The exception is the categories at each of the ends, which are unbounded.

sections, we use the domain of science in both assessments to exemplify this process. However, the points we make are generalizable to any domain in the assessments, and they serve as examples for our discussion on the choices made when generating PLDs from item maps.

5.5.1 IDs and PLDs in PISA and TIMSS

Figure 5.3 contains examples of one item from PISA and one item from TIMSS 8th grade. Both items belong somewhere above the middle of the scale in the item map, with percentages correct at 43% internationally. The PISA item is one of the items that define Level 4, while the quite similar TIMSS item anchors at the high international benchmark. The specific statement used to describe the TIMSS item is "Recognizes the major cause of tides" (Martin and Mullis 2012). In the international report from PISA, the item is presented as follows: "This is a multiple-choice item that requires students to be able to relate the rotation of the earth on its axis to the phenomenon of day and night and to distinguish this from the phenomenon of the seasons, which arises from the tilt of the axis of the earth as it revolves around the sun. All four alternatives given are scientifically correct" (OECD 2004, p. 289). In addition, the listing identifies the item as belonging to certain categories in the content and the procedural dimensions defined in the frameworks for the assessments[7].

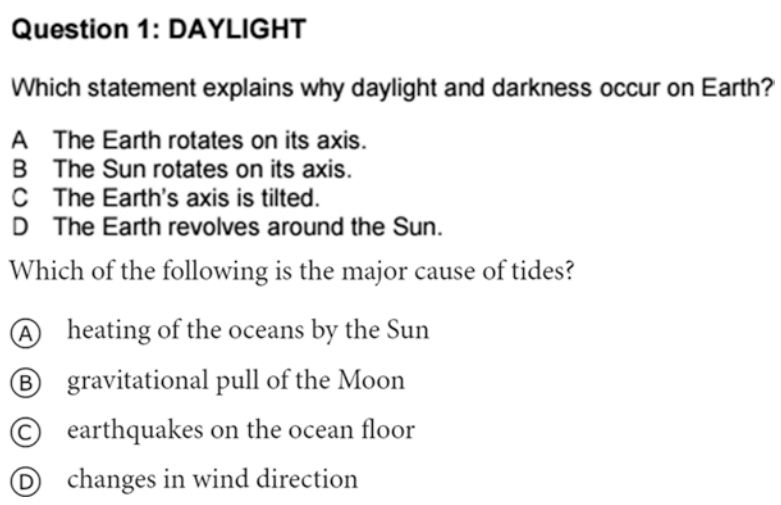

Fig. 5.3 Examples of items from PISA and TIMSS (the upper question is from PISA 2003, the lower question is from TIMSS 2011)

[7] In PISA, this aspect of the construct is defined by three competencies, and in TIMSS, by three cognitive domains.

These specific texts in the form of items are transformed into content-specific claims about what students with success on the items are able to do, or *IDs*, as we have coined them.

This stage involves some degree of generalization and removal from the original item-specific information. Finally, the full set of statements are reviewed, reduced, and synthesized into more overarching statements, PLDs, which express students' capabilities at discrete levels through a process involving consensus among subject matter experts.

There are some obvious differences in the two assessments' PLDs (see Table 5.1 for examples). The PLDs developed from TIMSS are longer and more detailed; furthermore, they refer more specifically to the content covered by the items. Also, the PLDs in TIMSS, given the more item-dependent language used, are not identical from one assessment to the next, while the statements used in PISA are almost identical over time. Although the PLDs in TIMSS also refer to what students are able to do with their knowledge ("compare," "contrast," etc.), the PLDs in PISA have a unique focus on such procedural aspects and include more generic competencies such as "reflect" and "communicate."

These differences reflect the divergent definitions and operationalizations of the domain of science in the two assessments. TIMSS has a framework with a high degree of content specification that is based on analyses of curricula in the participating countries. PISA is concerned with what students at the age of 15 are able to do in situations where an understanding of and about science (as a knowledge-generating process) is needed. The stimulus and items are therefore crafted to be less dependent upon very specific content knowledge.

Table 5.1 Examples of PLDs in PISA and TIMSS

PISA Level 4 (559–663)	TIMSS High (550)
"At Level 4, students can work effectively with situations and issues that may involve explicit phenomena requiring them to make inferences about the role of science or technology. They can select and integrate explanations from different disciplines of science or technology and link those explanations directly to aspects of life situations. Students at this level can reflect on their actions, and they can communicate decisions using scientific knowledge and evidence" (OECD 2007, p. 43)	"Students apply their knowledge and understanding of the sciences to explain phenomena in everyday and abstract contexts. Students demonstrate some understanding of plant and animal structure, life processes, life cycles, and reproduction. They also demonstrate some understanding of ecosystems and organisms' interactions with their environment, including understanding of human responses to outside conditions and activities. Students demonstrate understanding of some properties of matter, electricity and energy, and magnetic and gravitational forces and motion. They show some knowledge of the solar system, and of Earth's physical characteristics, processes, and resources. Students demonstrate elementary knowledge and skills related to scientific inquiry. They compare, contrast, and make simple inferences, and they provide brief descriptive responses combining knowledge of science concepts with information from both everyday and abstract contexts" (Martin et al. 2012, p. 83)

5.5.2 The Number and Nature of PLDs

The aim of assessments like PISA and TIMSS is to develop solid measures of students' proficiency in a few defined domains. At the outset, the items selected for the assessments are standalone, single observations of situations in which we can reasonably assume that students' overall ability on the measured trait is involved. However, single items are very unreliable observations of these abilities. The items involve unique content, make use of idiosyncratic language and representations, and various response formats. For a typical test, simple isolated right/wrong items have a point biserial correlation with the overall test score in the order of 0.3–0.4. In psychometric terms, this means that only 9–16% of the variance for an item can be seen as "true" variance related to the common trait being measured, whereas the major portion of the variance is residual variance (Olsen 2005). The obvious question regarding the PLDs developed from tests as part of a standard-setting process is to what degree we should include and rely on item-dependent information in the proficiency level descriptors. After all, the proficiencies we seek to describe are regarded as independent of the actual items included in the assessment.

Another important decision to be made is how many cut scores to identify and how to use them to assemble performance levels. In addition to reflecting the purpose of the PLDs, this decision is contingent upon the number of items available and how they are distributed across the scale. For many reasons, TIMSS has almost twice as many items in each domain as even the major domain in each PISA cycle. In this respect, TIMSS has a more favorable starting point for the process because more items mean more information to potentially include in the item maps. TIMSS has taken advantage of this by describing students' proficiency at or close to a few points or benchmarks on the scale. In this process, items with very similar difficulties are identified and clustered, which enables the development of PLDs by aggregating and synthesizing information across a set of data points with shared properties. This also allows for the development of PLDs that are well separated along the continuous scale. As a result, PLDs with a clear progression from one level to the next are produced. However, this method also leads to excluding items that do not meet the strict anchoring criteria; in fact, only half of the items fully satisfy the criteria. But, by including items that are almost anchoring, the number of items used to develop most of the PLDs in TIMSS is relatively large and should constitute robust data in the process. PISA, given its more limited amount of items, opts to describe intervals along the whole scale; thus, it includes all of the items in its process of extracting PLDs. Given this relatively low number of items, the decision to develop PLDs for six distinct levels on the scale seems rather ambitious.

We have not seen an overview of the number of items in the different levels in PISA, but assuming that the item difficulties resemble a normal distribution, we suggest that the number of items in the top and bottom levels is rather low. For the reading assessment developed for PISA 2009 and the mathematics assessment for PISA 2012, efforts were made to include a larger number of easy items. This was a well-reasoned improvement, given that the cut score between Level 1 and Level 2

receives policymakers' attention. Another issue related to the high number of PLDs in PISA is the risk that the progress in proficiency involved in advancing from one level to the next may not be that easy to grasp when reading the PLDs from low to high. For the same reason, Perie (2008) recommends using no more than four levels.

Thus far, we have established that the standard setting processes in PISA and TIMSS more or less follow the same principles: item maps are produced, IDs capturing both highly item-specific information and the more generic aspects involved in the construct are formed, and PLDs are extracted through an expert consensus process. Figure 5.4 illustrates this step in the standard setting process as a continuous scale ranging from completely item-specific statements to descriptions of generalized proficiencies. TIMSS has developed PLDs with a closer reference to the content of the items than PISA. PISA has developed PLDs with generic statements more closely resembling a theory of what constitutes progress in scientific literacy.

We argue that depending on the number of items at hand, the purpose of the assessment, and the intended use of the reported results, a decision should be made regarding where on this continuum it is possible and useful to target the PLDs. To the far left of this spectrum (see Fig. 5.4) are extremely item-specific PLDs, for instance in the form of a listing of all of the IDs. To the far right are very generic PLDs, for instance, in the form of simple labels such as low, intermediate, and advanced with only short and unspecific descriptions. The very item-specific information available in PLDs toward the left-hand side of the figure, could provide teachers with relevant subject matter information to be used in their formative practices. However, PLDs at this end are less robust in that they are more contingent on the actual items included in the test. Descriptors belonging to the right-hand side in Fig. 5.4 are less dependent upon the actual items included in the test and could for instance be used to communicate more generalized understandings of what constitutes performance on different levels in the construct being measured. Such PLDs could serve grading purposes and they could also potentially be used in assessments where learning progressions over longer time-spans are monitored.

Given the very high number of items available, we suggest that TIMSS should develop PLDs that are more generic and stable over time. After all, the assessment aims to report measures that are linked from one assessment to the next. For PISA, we suggest that more items are needed in order to develop PLDs in their current generic form. One possibility to remedy this situation is to create a new standard-setting process in which all available assessment material used since the first assessment in 2000 is assembled into one item map. With the 2015 assessment, all three domains in PISA have served as major domains twice, and pooling all of the item

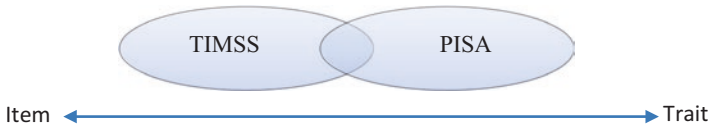

Fig. 5.4 Item versus trait specificity in TIMSS and PISA

information across assessments would significantly increase the number of IDs available for generating PLDs.

5.6 Possible Implications for the Norwegian Context

Other chapters in this book include more explicit descriptions of the current use of standard setting in the Nordic countries. Thus, here, we focus on suggestions for why and how standard setting where PLDs are developed for reporting purposes should be considered in the Norwegian context. With our partial knowledge of the situation in the other Nordic countries, we assume that this discussion is relevant for other Nordic countries. We first discuss some issues related to the national assessments before returning to issues related to the interpretation of grades in exams. Both types of assessments are reported using a limited number of reporting categories organized from lower levels to higher levels of achievement. In this sense, both assessments use standards established in an empirical setting.

5.6.1 PLDs in National Assessments

Explicitly formulated PLDs have already been created for the Norwegian national assessments[8]. The assessments are low stakes and are conducted at the beginning of 5th and 8th grade in the cross-curricular domains of reading, numeracy, and English. A description of how these PLDs were developed is, as far as we know, not publicly available. Without going into details of the nature of these PLDs, some similarities and some differences exist across the three domains. They all describe three levels for 5th grade and five levels for 8th grade. Originally, these cut scores were determined via specific percentiles. The PLDs in the reading domain resemble those used in PISA, while those developed for numeracy and English include more content-specific statements. Numeracy operates with overarching and generic PLDs in addition to a list of very content-specific statements.

We suggest that the methods applied for standard setting in the international assessments could be helpful in revising and document a transparent basis for the current PLDs. After initially using classical test theory and percentiles as the basis for reporting, all of the national assessments in Norway are now being developed within an IRT framework. Moreover, the assessments are now linked over years to support interpretations of trends. With these changes, new challenges and possibilities for standard setting have emerged. In the process of revising the standards we suggest that two issues should be emphasized. First, with the new scales developed to link performance over time, there is a need for robust descriptions at a more

[8]The actual descriptions are available in Norwegian from http://www.udir.no/Vurdering/Nasjonale-prover/

general level without reference to specific item content (also suggested for TIMSS). Furthermore, it is now possible and potentially very helpful to develop a joint item map for each assessment domain and each grade level in the national assessments, including the complete item material from several years of testing. This would help producing even more IDs, which would be particularly helpful for developing more robust PLDs for low- and high-performing students (in line with what we recommend for PISA). Furthermore, the progress from 5th to 8th grade, possibly extending to include 11th grade[9], should be explicitly modeled in the new PLDs. Standards with such a vertical scaling perspective are more challenging to develop because aligning PLDs across grade levels must be taken into account.

5.6.2 New Standard-Based Exams?

Given that the exams have multiple purposes, are high stakes for students, and are laborious and resource-intensive processes, it is unfortunate that the grading system appears to be unfixed and allows for inconsistencies and arbitrariness. A few examples supporting this claim can be found from official statistics reported in yearly national publications, e.g., The Norwegian Directorate for Education and Training (2014):

- Half of all pupils achieve lower written exam results than they do coursework grades in the same subject.
- The difference between coursework grades and exam grades varies systematically across schools.
- Even though the general descriptions of grades are the same for all subjects, the variation in average grades across subjects is large.
- Average grades, particularly for exams, vary over years

These observations illustrate that not all aspects of current grading practices are well understood. Assigning grades to students is defined as a judicial act, and these examples indicate a lack of transparency in current grading practices. Establishing more robust standards could be one helpful way to improve the situation.

However, standard setting in this situation is far more complex than setting cut scores and extracting PLDs. First, grading coursework typically includes evaluating products, not just assessments in the form of standard tests. Second, grades are formally defined to represent the degree to which the students have demonstrated mastery of the intended curriculum. In reading the curricular aims for a subject, it is quite evident that they are not formulated to reflect a unidimensional trait that lends itself to measurement on a single scale. Instead, most appear piecemeal with non-related descriptions of knowledge and processes that students should master.

A recent committee touched upon this issue in their series of white papers discussing the future of the Norwegian education system (NOU 2014:7, 2015:8). In

[9] Similar assessments are available for 11th grade, but they are not compulsory.

these reports, they recommend developing systems that support deep learning and learning progression. They do not explicitly state how learning progressions should be formulated or achieved, but in order to support progression, the formal curriculum needs revision. Care should be taken in reformulating curricular aims with a clear conceptual progress across grades. It is unreasonable to expect that subject matter expert groups working in isolation could formulate curriculum standards with such properties. Standard setting procedures, including collecting and analyzing empirical data in some form or another, are needed to support this process.

We do not claim that all of the issues related to the complexity of grading students in school may be fixed by simply performing one or several standard-setting procedures. However, as exams are already very systematic and large-scale logistic operations, it is possible to collect data and develop item maps as described above. This could constitute the first small step toward a more robust foundation for grading in the Norwegian school system.

References

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191–204.

Beaton, A. E., & Zwick, R. (1992). Overview of the national assessment of educational progress. *Journal of Educational and Behavioral Statistics, 17*(2),95–109.doi:10.3102/10769986017002095.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93–106. doi:10.1111/j.17453984.1993.tb01068.x.

Cizek, G. J. (Ed.). (2012). *Setting performance standards foundations, methods, and innovations*. New York/London: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting : A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 79–106). New York/London: Routledge.

Gonzalez, E. J., Galia, J., Arora, A., Erberber, E., & Diaconu, D. (2004). Reporting student achievement in mathematics and science. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 274–307). Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Gregory, K. D., & Mullis, I. V. S. (2000). Describing international benchmarks of student achievement. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 265–276). Chestnut Hill: International Study Center, Boston College.

Huynh, H. (2009). Psychometric aspects of item mapping for criterion referenced interpretation and bookmark standard setting. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 148–159). Maple Grove: JAM Press.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring.* (PhD), Boston College, Boston.

Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 International results in science*. Chestnut Hill: TIMSS & PIRLS International Study Center.

Mullis, I. V. S. (2012). Using scale anchoring to interpret the TIMSS and PIRLS 2011 achievement scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschof, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

NOU. (2014:7). *Elevenes læring i fremtidens skole. Et kunnskapsgrunnlag [Pupils' learning in the school for the future. A knowledgebase]*. https://nettsteder.regjeringen.no/fremtidensskole/

NOU. (2015:8). *Fremtidens skole. Fornyelse av fag og kompetanser [A school for the future: Renewing school subjects and competencies]*. https://nettsteder.regjeringen.no/fremtidensskole/

OECD. (2001). *Knowledge and skills for life. First results from PISA 2000*. Paris: OECD Publications.

OECD. (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: OECD Publications.

OECD. (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: OECD Publishing.

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Technical report*. Paris: OECD Publishing.

Olsen, R. V. (2005). *Achievement tests from an item perspective. An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science*. Oslo: Unipub forlag.

Perie, M. (2008). A guide to understanding and developing performance level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15–29. doi:10.1111/j.1745-3992.2008.00135.x.

Smith Jr., E. V., & Stone, G. E. (Eds.). (2009). *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models*. Maple Grove: JAM Press.

The Norwegian Directorate for Education and Training. (2014). *The education mirror 2014: Facts and analysis of kindergarten, primary and secondary education in Norway*. Oslo: The Norwegian Directorate for Education and Training http://www.udir.no/globalassets/upload/rapporter/educationmirror/the-educationmirror_english.pdf.

Zieky, M. J., Perie, M., & Livingstone, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service.