Setting benchmarks on the UIS reporting scales

Discussion Paper for GAML Task Force 4.1
June 2017

The ACER Centre for Global Education Monitoring supports the monitoring of educational outcomes worldwide, holding the view that the systematic and strategic collection of data on education outcomes, and factors related to those outcomes, is required to inform high quality policy aimed at improving educational progress for all learners.

ACER

Contents

Introduction

This paper has been prepared to inform a discussion about the setting of benchmarks on the UIS Reporting Scales. Whilst prepared with a view to the reporting of indicator 4.1.1, the discussion and issues raised will be relevant to all indicators related to the measurement of learning outcomes

Indicator 4.1.1 refers to the percentage of children/young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics.

In order to monitor progress against this indicator, within and across countries, it is necessary to establish uniformly applicable definitions of the following features of the indicator:

1. What is reading;
2. What is mathematics;
3. What it means to be in grades 2/3;
4. What is the end of primary schooling;
5. What is the end of lower secondary schooling;
6. What is minimum proficiency in (i) reading and (ii) mathematics for each of the three target levels mentioned in points 3–5.

The definitions of reading and mathematics will be established via the Global Common Framework for Reference and the UIS Reporting Scales. This paper deals with issues (3) to (6) on the assumption that a progression of developing knowledge, skills and understandings in each of reading and mathematics has been developed and is widely accepted.

In the following we outline the issues associated with establishing the necessary definitions and present some discussion of processes that can lead to the definitions being established.

Throughout this document we will refer to the location on the UIS Reporting Scale that represent minimum proficiency as *benchmarks*.

Defining the three points of schooling that relate to Indicator 4.1.1

Rationale

In order to monitor the achievement of goals within a given education system, it is important that the definitions used for each point in schooling are used consistently within the system over time.

For example, a robust measure of progress against the indicator would be to compare the proportion of Grade 3 students meeting a given benchmark in one year of data collection to the proportion of Grade 3 students meeting the same benchmark in another year of data collection. Progress against the indicator would not be reliably determined if the grade at which achievement was measured changed from year to year, nor if the benchmark to which achievement was compared changed from year to year.

For the purpose of monitoring progress *within a given education system*, the choice of whether Grade 2 or Grade 3 is selected as the target year level for testing is only relevant

when considering the appropriateness of the benchmarks (i.e. the level of difficulty and content) against which student achievement is being compared.

In contrast, the selection of the target year level for testing *across education systems* can be highly sensitive. In order to make meaningful comparisons across education systems, it is important that developmental context (such as the number of years of schooling and the age/maturation of the students) is as consistent as possible. This poses a number of practical and political issues associated with the different structures of schooling across systems and the political imperatives that guide policy development within and across countries. In short, the differences that exist in the structure of formal schooling (such as the age at which students enter formal schooling or the match between curriculum content and student grade/age/development) are sufficiently large across countries that they can challenge the credibility of cross-national comparisons. It is therefore necessary to have unambiguous and unequivocal guidelines around the selection of the target assessment grades at each of the specified levels under Indicator 4.1.1.

Age versus grade-based definitions

There are two ways in which the target assessment levels can be identified. One is according to their grade level in education (grade-based) and the other is according to their age (age-based). The indicator 4.1.1 refers to 'grades 2/3', 'end of primary' and 'end of lower secondary' and this probably suggests a preference for a grade based definition. The language, however, is not specific and discussion will be required to prepare an operational definition. In addition, to be applicable for all children, including those not in school, aspects of an age may be required in an operational definition.

Both grade-based and age-based approaches are currently used in cross-national assessments. For example, PISA uses an age-based sample (15 year olds) whereas TIMSS (Grades 4 and 8), PIRLS (Grade 4), ICCS (Grade 8) and ICILS (Grade 8) use a grade-based sample.[1]

Determining whether to define the population of interest by age or grade depends on the purpose of the data collection exercise and the nature of the expected comparisons. It is also influenced by practical issues. An age-based population definition provides the possibility for truer maturational-based comparisons, but it also makes it more difficult to draw conclusions about the influence of educational experience on student outcomes, because the students being compared are typically drawn from a range of grade levels. Age-based population definitions can also be more expensive and difficult to implement than grade-based population definitions, as typically students must be sampled from across classes within schools rather sampled as intact classes. While a grade-based population definition suggests a truer comparison of the outcomes associated with specific curriculum exposure (ie up to and including the grade of assessment), in the context of cross-national assessment the data are susceptible to variations associated with policies affecting the school starting age and grade retention, which in turn can confound interpretation of the data.

Ultimately both approaches are sensitive to policy differences across countries in terms of the educational experiences provided to students. The key is that the analysis and reporting of

[1] In practice, PISA, TIMSS, PIRLS and ICCS rely on population definitions that refer to both age and grade information, but in the case of PISA the age element of the definition has primacy and the grade element simply establishes the lower bound for the grade level of the target population, whereas in the IEA studies the grade element of the definition has primacy, and the age element places a lower bound on the age of the target population.

student outcomes clearly take into account the strengths and limitations associated with whichever method is used.

The grade-based language used in Indicator 4.1.1 ('Grade 2/3', 'the end of primary' and 'end of lower secondary') will need national to be applied in a sufficiently standardised form to permit reasonable across country comparisons. The International Standard Classification of Education (ISCED) is 'the reference classification for organizing education programmes and related qualifications by education levels and fields' (OECD, European Union, UNESCO Institute for Statistics, 2015, p. 9) against which grade-based selection of the relevant sample can be referenced. This system 'employs proxy criteria to classify a given educational program that help classify a given education programme to the appropriate ISCED level' (OECD, European Union, UNESCO Institute for Statistics, 2015, p. 14) in order to account for 'the absence of direct measures to classify educational content (OECD, European Union, UNESCO Institute for Statistics, 2015, p. 14).

Table 1 shows a classification of the three target levels specified in Indicator 4.1.1 against the ISCED levels.

Table 1: ISCED classification of years of schools in indicator 4.1.1

Year/Grade indicated in Indicator 4.1.1	ISCED Level
Year 2/3	Mid-ISCED Level 1
End of primary	End of ISCED Level 1
End of lower secondary	End of ISCED Level 2

In the context of monitoring Indicator 4.1.1, the ISCED levels are useful for the purpose of interpreting local country-specific grade definitions against a standardised international framework. However, they do not specify the actual grade level that would be monitored within countries and may represent a large variation in the age of children. For example, in Australia Year 2/3 would correspond to 8-9 years of age, whereas in South Africa it would correspond to 10-11 years of age. So across the world Year 2/3 could allow an age range of 3 years (~30%) and at least a year in schooling (~25-50%).

If a grade-based selection mechanism is preferred, the grades that correspond to the target measurement points, as specified by ISCED levels, for monitoring of Indicator 4.1.1 will need to be specified at the country level, with each country suggesting which grades best correspond to the target measurement points. Further an out-of-school equivalent may well be required. This process would require some form of central adjudication to maximise the viability of corresponding cross-national comparisons. While in many cases this may be unproblematic, there may be cases in which this is a politicised and sensitive process. The system of adjudication would need to be sufficiently sensitive and nimble to negotiate the challenges that may be presented by some countries nomination of target grades.

Recommendation 1 – Use ISCED to in providing a cross-nationally standardised way of referring to the measurement points in Indicator 4.1.1. However, the in-precision in terms of years of schooling and applicability to out of school cohorts will need to be considered.

Recommendation 2 – Countries' specifications for the target grades that correspond to the measurement points in indicator 4.1.1 will need to be adjudicated against an agreed set of criteria.

Establishing an appropriate level of specificity for the grade levels

Year 2/3

Year 2/3 refers to two grade levels that are in a period of schooling where learning growth in one year can be substantial. Whilst Year 2 and 3 are both in the middle of ISCED Level 1 the formulation Year 2/3 does leave open the possibility of considerable variation. For the purpose of establishing achievement standards and monitoring outcomes against them, it would be preferable to adopt a referent year level that did not permit multiple interpretations. Establishing a clear definition of the referent year level would support specification of the standards (in each of reading and mathematics) relative to expected learning at that point in schooling, and also reduce likely variation across countries in the selection of the target students. Nominating one year as the target would not reduce the applicability of measurement against the standard across countries, as, subject to adjudication, individual countries would still be responsible for selecting the grade level that was 'equivalent' to the measurement point in the indicator.

As discussed previously, under such a model, the flexibility to select the target grade that maximises cross-national standardisation can come through allowing countries to suggest their target grade according to the structures of the education systems rather than by using a very broad definition of the target grade.

End of primary and end of lower secondary

Whilst the end of primary is equivalent to the end of ISCED Level 1 and end of lower secondary is equivalent to the end of ISCED Level 2 it may still be that these definitions are not sufficiently precise, for example, some countries have longer and shorter ISCED 100 and 200 programs which mean the "end" of primary could reflect a different number of years of input. Out-of-school equivalents may well still be required.

Recommendation 3 – Adopt more precise interpretations (ie than the current 'Grade 2/3', 'the end of primary' and 'end of lower secondary'*) for the target groups and consider the implications for an out-of-school equivalent.*

Defining *minimum proficiency* in each of reading and mathematics

In the context of Indicator 4.1.1, it is assumed that, for each of reading and mathematics, a standard that represents *minimum proficiency* will be established for each of the three target year levels such that the percentage of students meeting and not meeting the standard can be measured.

As such, there is need to establish a substantive definition of minimum proficiency and a related operational definition of minimum proficiency against which student achievement can be measured. In the following sub-sections we will first discuss the issues associated with establishing definitions of minimum proficiency, followed by some discussion of standards-setting methods that could lead to the establishment of operational definitions of minimum proficiency. In this discussion we assume, for convenience, that the standards will be established against the UIS reporting scales. There is first a question of whether a single definition of 'minimum proficiency' is to be set or whether the target requires a different

'minimum proficiency' for each year level? Further if three different 'minimum proficiency' (ie located at different points on a scale), it is feasible that the conceptual foundation for 'minimum proficiency' could vary across the three target year levels. For example, a definition of 'minimum proficiency' at Year 2/3 could relate to demonstration of basic functional proficiency in reading or mathematics, whereas the definition at end of lower secondary could be specified in terms of preparedness for further (upper secondary or vocational) education. Such flexibility would allow for the definitions of minimum proficiency to be tailored to best fit the targe grade levels. However, using different definitions across the levels could lead to confusion about which the standard applies in which context and questions about why different definitions have been used. On balance, it seems wise to adopt a single descriptive definition of the standard (i.e. minimal proficiency) for all three grade levels.

Recommendation 4 – Adopt a single descriptive definition of the standard (i.e. minimal proficiency) for all three grade levels.

Norm- versus curriculum- or content-based standards

One of the first decisions that needs to made is whether the standards are to be established against populations norms or against substantive content that relates to curriculum outcomes or otherwise described standards. If established against population norms the benchmark would be framed as, for example, the scale score representing the 10th per centile in a reference year of students in one or more 'reference' countries. In contrast, if established against content, then the benchmark would be a scale score that represents a given set of described achievement outcomes.

Content-based standards are established for the purpose of determining whether students have achieved a specified set of knowledge, skills and understandings. In contrast, norm-referenced standards are based on the distinction between 'higher' and 'lower' achieving students, without the necessary expectation that the knowledge, skills and understandings that differentiate between 'higher' and 'lower' achievement need be described in detail, or indeed at all. Given the broader purpose of monitoring SDG 4, we recommend establishing content-based standards that are informed by and can be mapped to local curricula and relevant national and international standards.

Establishing content-based standards at each of the three specified points in schooling will be transparent and anchor the standards in the substance of each learning area. This would also remove the challenges associated with choosing the referent population (such as 'students in countries X, Y and Z') that would be required for establishing normative standards. In addition to the political sensitivities associated with determining the norm-referents, there may be challenges for new countries joining the monitoring program in accepting pre-determined norm-references that they feel not to be relevant or applicable to their own circumstances. In contrast, even though there may be arguments about the substance of curriculum- or content-based standards, these arguments can focus on the substance of the standards and consequently are easier to communicate and more transparent.

It is important to note, however, that content-based standards will need to be informed by normative data where such data exist. For example, it may be regarded as unrealistic or impractical if a minimum standard is set against a curriculum and only 40% of students achieve the standard. However, in the context of establishing standards against which to monitor SDG 4, one significant challenge to the use of normative data will be presented by the variation in achievement that exists across (and within) countries. For example, it is feasible that a minimum standard that is set against curriculum outcomes may be achieved by 90% of students in a high achieving country and by 50% of students in a low achieving

country. Where a uniform standard is set and applied across countries, it is inevitable that there will be variation in achievement of the standard across countries. Under such circumstances, it is communication around what is valued in a monitoring program (for example, a focus on improvements within countries over time rather than necessarily on cross-country comparisons) that can engender support for and trust in the program.

Recommendation 5 – Though benchmarks should be content-referenced, their establishment should be informed by normative data where such data exist.

Using curricula to inform the standards

It will be necessary to determine how the standards reflect a notion of 'minimum proficiency' that is applicable across countries.

If a curricula are used to inform the standards, then the essence of 'minimum' reflects a sense of 'minimum proficiency following X years' exposure to curriculum'. Such a focus implies that a form of curriculum audit would be conducted across countries so that common and reasonable expectations of minimum learning based on exposure to the curricula could be determined.

In addition, common agreement would need to be reached on the notion of 'minimum'. For example, this may refer to the 'minimum level of achievement that would allow a student to likely succeed in their next year of learning without the need for specialist intervention', or 'the minimum level of achievement that would allow the student to function in and outside of the classroom', or 'the minimum standard to expect from a student who has been at school for X years'.

Recommendation 6 – Agreement will need to be reached about the interpretation of the expression 'minimum proficiency' for each measurement point for each domain. A process for achieving this agreement is required.

It would be possible to establish an agreed set of skills, knowledge and understandings that reflect 'minimum proficiency' independent of curricula but with a broader reference to real-world functioning. However, school-based achievement as a preparation for later life is a strong focus in primary and early secondary schooling (the target levels for monitoring Indicator 4.1.1) and curricula are typically more consistent across countries at these lower levels of schooling. We recommend ensuring that the standards are informed by curriculum in a process that involves collecting curriculum information from countries, comparing it to the standards, and further refining the standards based on the findings of that comparison – that is mapping national information to the reporting scale. This will be an iterative process of collecting curriculum information from countries, comparing it to the standards and further refining the standards to ensure that the curricula appropriately inform the standards. It is important to note that, while the UIS reporting scales can provide the substance of the standards, it is the process of achieving satisfactory substantive definitions for the benchmarks, and later establishing the positions of the benchmarks on the scale, that will need to be informed by reasonable expectations of student learning against a range of national curricula. Further to this, consideration should be given to how national and international benchmarks compare to the benchmarks established for monitoring Indicator 4.1.1. Many countries will already have benchmarks that are used for national monitoring and reporting, and any benchmarks established for monitoring against indicator 4.1.1 should ideally be complementary to these national benchmarks, or at least not acting in opposition to them.

Similarly, the standards established for indicator 4.1.1 should not contradict or act in

opposition to any established standards from existing international or regional assessments. In addition to this, there are benchmarks established in TIMSS for Grades 4 and 8 in mathematics, PIRLS in Grade 4 for reading, and other regional and supranational indicators that may be considered when establishing the standards. In effect, any newly established benchmarks should not, without good reason, contradict or act in opposition to existing national or cross-national benchmarks.

Recommendation 7 – The establishment of the benchmarks on the UIS scales will need to be:

- *informed by curricula from a variety of countries.*
- *An iterative process*
- *Consistent with existing national and international standards.*

Should the standard be a point or range on a scale

A content-based operationalisation of the standard at each grade level will result in each benchmark having a set of knowledge, skills and understanding that are demonstrated by students who have met (or exceeded) the benchmark. These descriptions will likely be synthetic generalisations of the knowledge, skills and understandings that students demonstrate when completing tasks that are in the same range of difficulty on the scale as the benchmark. There are two possible approaches to conceptualising the operational definitions of the benchmarks.

The first approach is to have each benchmark represented by a single point on the scale and, have the description of the benchmark be as accurate as possible a synthetic description of the characteristics of achievement at that point on the scale. The main advantage of this approach is that the benchmark is, by design, unambiguous. One disadvantage is that the description of the benchmark as a point suggests an unrealistic level of precision that is not consistent with the uncertainty that will be inevitable in the alignment process that maps nation data to the UIS Reporting Scale.

The second approach is to have each benchmark represented as achievement of a ‘level’ on a scale, defined by a score range and described by a synthetic description of the knowledge, skills and understandings that describe achievement within that range. An arguable disadvantage of this approach is that at face value there is not a clear single point on the scale that represents the benchmark. However, in reality the upper limit of the defined level is the point on the scale that represents the benchmark for the purpose of monitoring achievement. The main advantage of this approach is that stakeholders typically feel more comfortable interpreting a ‘level’ as a target range of achievement, rather than focusing on a single point on the scale. This approach may also lead to greater stakeholder ease and comfort in the process of establishing the benchmarks as there is naturally a sense that in working with a range, rather than with a single point, there is greater flexibility in the process. Ultimately it is a matter of stakeholder perception that drives the selection of the benchmark as a point or as a range on the scale.

Recommendation 8 – Consideration will need to be given to whether the benchmarks should be points or ranges on the UIS scales.

Setting the benchmarks

Introduction

In this sub-section we discuss the fundamentals of a process of setting performance benchmarks in the context of monitoring SDG 4. This process focuses on the use of expert judgement to define points on a scale such as the UIS reporting scale that represent a generic definition of 'minimum proficiency'. The process is assumed to follow the same steps for each of reading and mathematics, and if required, for each of the three target measurement points. The same principles apply to the process regardless of whether each benchmark is established as a point on the scale or a range on the scale. In the research and methods literature procedures for setting benchmarks are usually referred to as standard-setting methods.

Steps in using standard setting to set benchmarks

There is no fixed set of steps to be conducted in standards setting. Hambleton (1998) described a generic set of steps in standards-setting exercises that are common to the process of setting performance standards, regardless of the specific approach or procedure adopted, and these have been adapted by others such as Cizek (2006) and Hambleton (2012). An adapted set of these steps is provided in Table 2.

Table 2: Generic steps in setting performance standards

Step	Description
1	Choose a standard-setting procedure: prepare training materials and standard-setting meeting agenda.
2	Select a large and representative panel of experts including regional and international assessment network plus country representative.
3	Prepare descriptions of the referent candidate or group and of the performance categories.
4	Train participants to use the standard-setting method.
5	Compile item judgements/ratings from experts and summarise outcomes to provide feedback on ratings.
6	Facilitate a discussion amongst the experts based on the feedback from the rating exercise.
7	Provide experts the chance to revise their ratings on the basis of the feedback and discussion; this may include repeating Steps 5 and 6.
8	Ask experts to review/reconfirm their ratings to determine a 'final' recommended standard.
9	Conduct an evaluation of the process with the participants to confirm that they are satisfied with and have confidence in the process.
10	Assemble documentation of the process and other evidence that may have a bearing on the validity of the resultant standards.

Steps 1, 2 and 3 are the preparatory steps that frame the standards-setting process. These will

be discussed in some detail. Steps 4–10 are presented in Table 1 for completeness, and operational details of them would be established as the logistical parameters of the monitoring program are developed, but these steps do not warrant in-depth further discussion at this point in time.

Step 1: Choosing a standards-setting method

There are many different standards-setting methods which can be classified into two broad categories – test-centred methods and examinee-centred methods.

Test-centred methods

In test-centred methods, judges review the items/tasks completed by students and set the standards relative to their judgement of the difficulty of the items/tasks. Test-centred methods focus on the content of test materials (which may include achievement scales and frameworks as well as test items) in establishing consensus judgements of the standard. The Angoff (and variations) and Bookmark methods are examples of test-centred methods.

Examinee-centred methods

In examinee-centred methods, judges review samples of work completed by candidates to establish categories of performance and then establish cut-scores (standards) that correspond to the distinctions between the categories. Examples of examinee-centred methods are The Paper Selection, Body of Work, Contrasting-Groups and Borderline-Group methods.

Adopting a multiple-method approach

While the division of methods presented above is common in discussions of standards setting, in practice multiple-methods are frequently used, and the implementation of any given method may be adapted to include aspects of other methods. In fact, there remains no agreed best method for setting standards (see, for example Zieky, 2001 or Linn, 2003). There is, however, agreement that, wherever possible, it is wise to consider the outcomes of *more than one method* and any additional relevant statistical information when establishing standards (Linn, 2003, Jaeger, 1989).

As discussed previously we are assuming that the standards themselves will be content-based in the sense that as they will be developed with reference to curriculum expectations, relevant existing national and cross-national standards and test items. However, setting the standards for monitoring Indicator 4.1.1. is an unusual context (when compared to the majority of standards-setting exercises that are highly localised) in that the candidature is so broad (cross-national) and that data are collected against the standard may also vary across contexts.

The use of more than one method provides a form of internal validation that helps account for a method-effect from biasing the experts' judgements. It will also allow for greatest flexibility in accommodating the logistics of working with a broad range of experts from a range of countries.

Furthermore, providing judges access to performance data at points in the process can be a way of moderating judges' expectations. For example, Linn (2003) states:

> *Assuring that judges on standard setting panels understand the context in which the standards will be used is a minimal requirement for obtaining reasonable performance standards. Normative information needs to be made part of the process for judges to anchor their absolute judgments with an understanding of current levels of performance of students and likely consequences. As Zieky (2001) has noted,*

considering both absolute and normative information "in setting a cut-score can help avoid the establishment of unreasonably high or low values" (p. 38).

Recommendation 9 The approach adopted to set the benchmarks should be primarily test-centred but include more than one method and, where possible, draw on performance data to support judges' decisions.

It is too early at this stage to suggest an exact procedure, given how little is currently known about the logistics of running the activities in the context the UIS reporting scales (such as how many experts are likely to be used and whether or not they could meet face-to-face or whether a remote activity would need to be conducted). One practical blended method uses a Modified Angoff (Yes/No) (see, for example Cizek, 2006 or Kellow and Wilson, 2008), followed by review of normative data and then a Bookmark method procedure (see, for example, Mitzel et al, 2001, or Kellow & Wilson, 2008) to set standards across a range of different contexts. Without presupposing that this model would be used, it provides an example of a multiple-method set of activities that incorporates empirical assessment data.

As the standards-setting method is refined and confirmed over time it is important to keep in mind the observation of Hambleton et al (2012) that 'the choice of method is often a much less significant factor that the way in which the method is actually implemented'.

Selecting a large and representative panel of experts

The common element to all standards-setting approaches is that they 'invoke the judgement of *experts* in the content area of interest' and that 'these individuals possess substantive content knowledge as well as intimate familiarity with the target population' (Kellow & Wilson, 2008).

The greatest influence on the quality of outcomes from the standards-setting process is the level of expertise (in the discipline and with knowledge of the target population) of the judges. In the case of monitoring SDG 4, there is an additional political imperative that as large a range of countries as is practical should be involved in the process of establishing the standards.

The experts should be selected on the basis of their expertise in education in the relevant learning domains (reading or mathematics) and levels of schooling. It is important to remember that for establishing the standards relevant to Indicator 4.1.1 there will need to be six panels of experts. It is possible that some experts may participate in more than one panel (such as being on more than one panel in reading or mathematics) but it may be impractical for any given expert to participate in more than two panels.

It is not anticipated that the expert participants are 'representatives' of their countries, rather they should be nominated and selected on the basis of their expertise. However, we do assume that the members will be drawn from a large range of countries and that part of their contribution will be their expert knowledge of curricula and pedagogy in their countries.

Without yet being able to specify full details of the standards-setting process, we suggest that the process should be set up to support a large panel of experts. If a given panel becomes too large to function efficiently then it can be split into two or more smaller panels and their judgements later combined.

Recommendation 10 – The establishment of the benchmarks will require the establishment of a panel of experts. The panel should be:

- *Selected from national nominees*
- *Have a high level of expertise in education in the relevant learning domain*
- *Large*
- *Well resourced*

Prepare descriptions of the referent candidate or group and of the performance categories

At the end of the previous sub-section, we included a set of preparatory steps that need to be undertaken before the standards-setting exercise can be undertaken. In particular it will be necessary to have clear definitions of the three target levels of students and of the definition of minimum proficiency that will be considered by experts when they establish the standards.

Regardless of the specific standards-setting methodology that is adopted, all experts will need to develop a clear (and ideally consistent across experts) schema of a student demonstrating minimum proficiency in the relevant target grade level and domain. The quality of the schemas will depend on the quality of the definitions provided to the judges, the quality of their training in consolidating their schemas and of course the judges' expert knowledge and understanding of the relevant subject and target grade students.

Recommendation 11 – Expert panels will need to be supported to develop and consolidate clear schema of what minimal proficiency in the domain looks like at the relevant measurement points. Support comes through providing them with high-quality training and clear, unambiguous definitions of key terms.

Outcomes of the standards-setting exercise

The ultimate aim of the standards-setting exercise is to provide advice on the points on the UIS reporting scale that represent minimum proficiency in reading and mathematics at each of the three target grade levels. We use the term *advise* deliberately to indicate that the process of confirming the standards will most likely require further consultation with stakeholders. Stakeholder engagement and consultation will be critical to the success of the monitoring program and, as part of clarifying the parameters of the monitoring program, it will be necessary to establish the status of the standards-setting exercise such that participants and stakeholders understand that the recommendations from the standards-setting exercise will require review and confirmation. A possible approach is to implement methods in which the outcomes of the standards-setting exercise are that each standard is represented by a recommended 'acceptable range' within which the standard lies. In effect, each expert panel determines the lower and upper limits of the range (i.e. the points on the scale below and above which the panel agrees the standard should not be located). This allows for a degree of flexibility in the consultation that leads to confirmation of the standards.

Recommendation 12 – The benchmarks set by the expert panels should be submitted to broader stakeholder consultation before finalisation

References

Cizek, G.J. (2006). Standard Setting. In Downing, S. & Haladyna, T. (Eds.). *Handbook of Test Development (pp225-258).* Mahwah: Lawrence Erlbaum Associates.

Hambleton, R. M. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In Hansche, L. N. (Ed.). *Handbook for the development of performance standards* (pp. 87-114). Washington, DC: Council of Chief State School Officers.

Hambleton, R. M., Pitoniak M.J. & Copella, J. M. (2012): Essential Steps in Setting Performance Standards on Educational Tests and Strategies for Assessing the Reliability of Results. In Cizek (Ed.) *Setting Performance Standards Foundations, Methods and Innovations.* (pp 93-125) New York: Routledge.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Kellow, J & Wilson, L. (2008) *Setting Standards and Establishing Cut Scores on Criterion-Referenced Assessments Some Technical and Practical Considerations*. In Osborne, J (Ed.) Best Practices in Quantitative Methods (pp 14-28). Thousand Oaks: SAGE Publications, Inc.

Linn, R. (2003). *Performance standards: Utility for different uses of assessments*. Education Policy Analysis Archives, 2003, 11 (31).

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). Majwah, NJ: Lawrence Erlbaum Associates, Inc.

OECD, European Union, UNESCO Institute for Statistics (2015), *ISCED 2011 Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications*, OECD Publishing. http://dx.doi.org/10.1787/9789264228368-en. Creative Commons Attribution CC BY-NC-ND 3.0 IGO

Zieky, M. J. (2001). So much has changed: How the setting of cut-scores has evolved since the 1980s. In G. J Cizek (Ed.) *Setting performance standards: Concepts, methods and perspectives* (pp. 19-51). Mahwah, NJ: Lawrence Erlbaum.