Large-Scale Learning Assessments
A Handbook for the Indian Context
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of exhibits</td>
<td>6</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>8</td>
</tr>
<tr>
<td>Foreword</td>
<td>9</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>10</td>
</tr>
<tr>
<td>Bibliography</td>
<td>11</td>
</tr>
</tbody>
</table>

### 01 Introduction
1. The assessment context in India | 17
2. This handbook | 20
Glossary | 23
References | 24

### 02 Policy goals and issues
1. Education priorities | 25
2. The purpose of assessment | 27
3. Major stakeholders | 28
4. Project steering group | 30
5. Scope of the assessment | 31
6. Design elements of the assessment | 32
7. Considerations for policy implications and issues | 35
8. Summary | 37
Checklist: Policy goals and issues | 40
Glossary | 41
References | 42

### 03 Project team and infrastructure
1. Assessment centre | 43
2. Budget and source of funding | 44
3. Team organisation | 45
4. Core team management | 46
5. Capacity development | 47
6. Outsourcing | 51
7. Physical infrastructure | 52
8. Summary | 53
Checklist: Project team and infrastructure | 55
Glossary | 56
## 8 Test and questionnaire design

8.1 Test design 133
8.2 Questionnaire design 140
8.3 Summary 146

Checklist: Test design 147
Checklist: Questionnaire design 148

Glossary 149
References 151

## 9 Sample design

9.1 Purpose of sampling 155
9.2 Defining the target population 156
9.3 Defining the sampling frame 158
9.4 Stratification 159
9.5 Multi-stage sampling 160
9.6 Determination of sample size 162
9.7 Field trial sample design 164
9.8 Replacement protocols 165
9.9 Within-school sampling 166
9.10 Response rates 168
9.11 Sampling weights 170
9.12 Summary 171

Checklist: Sample design 172

Glossary 174
References 175

## 10 Materials production and field operations

10.1 Preparatory stage 179
10.2 Implementation stage 188
10.3 Post implementation stage 190
10.4 Summary 192

Checklist: Materials production and field operations 193

Glossary 195
References 196
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Data management</td>
<td>197</td>
</tr>
<tr>
<td></td>
<td>11.1 Codebook design</td>
<td>199</td>
</tr>
<tr>
<td></td>
<td>11.2 Data management software</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>11.3 Physical materials handling</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td>11.4 Data entry protocols</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>11.5 Data management training</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>11.6 Data cleaning, verification and validation</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>11.7 Database construction</td>
<td>210</td>
</tr>
<tr>
<td></td>
<td>11.8 Backup procedures</td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>11.9 Database security</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>11.10 Data transfer protocols</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>11.11 Quality control</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>11.12 Summary</td>
<td>215</td>
</tr>
<tr>
<td></td>
<td>Checklist: Data management</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>Glossary</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>218</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Scaling methodology</td>
<td>219</td>
</tr>
<tr>
<td></td>
<td>12.1 Scaling models</td>
<td>221</td>
</tr>
<tr>
<td></td>
<td>12.2 Methods for linking design</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>12.3 Test equating</td>
<td>227</td>
</tr>
<tr>
<td></td>
<td>12.4 Calibration</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>12.5 Estimating student ability</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td>12.6 Scale development</td>
<td>237</td>
</tr>
<tr>
<td></td>
<td>12.7 Summary</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>Checklist: Scaling methodology</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>Glossary</td>
<td>241</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>242</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Analysis for reporting</td>
<td>245</td>
</tr>
<tr>
<td></td>
<td>13.1 General methodology</td>
<td>247</td>
</tr>
<tr>
<td></td>
<td>13.2 Relationship between achievement and background variables</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>13.3 Sample weighting</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>13.4 Standard error and replication techniques</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td>13.5 Regression analysis</td>
<td>254</td>
</tr>
<tr>
<td></td>
<td>13.6 Multilevel analysis</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>13.7 Summary</td>
<td>259</td>
</tr>
<tr>
<td></td>
<td>Checklist: Analysis for reporting</td>
<td>260</td>
</tr>
<tr>
<td></td>
<td>Glossary</td>
<td>261</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>262</td>
</tr>
</tbody>
</table>
# Reporting and dissemination

## 14.1 Understanding the results

## 14.2 Dissemination strategy

## 14.3 Report writing

## 14.4 Facilitating policy impacts

## 14.5 Policy impacts

## 14.6 Summary

Checklist: Reporting and dissemination

Glossary

References

Contributors
## List of Exhibits

<table>
<thead>
<tr>
<th>Exhibit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhibit 1</td>
<td>13 key areas of a robust large-scale assessment program</td>
</tr>
<tr>
<td>Exhibit 2</td>
<td>Assessment purposes and study design</td>
</tr>
<tr>
<td>Exhibit 3</td>
<td>Example of an assessment team organisational chart</td>
</tr>
<tr>
<td>Exhibit 4</td>
<td>Statement of NAS Class X assessment purpose</td>
</tr>
<tr>
<td>Exhibit 5</td>
<td>Example of generic skills classification, NAS Class X assessment framework</td>
</tr>
<tr>
<td>Exhibit 6</td>
<td>Percentage distribution of text types in NAEP at three grade levels</td>
</tr>
<tr>
<td>Exhibit 7</td>
<td>Target ranges for processes (skills) from a numeracy assessment framework for initial teacher education students</td>
</tr>
<tr>
<td>Exhibit 8</td>
<td>Structure of the NAEP Writing Assessment Framework, 2007</td>
</tr>
<tr>
<td>Exhibit 9</td>
<td>Structure of the PIRLS assessment framework, 2011</td>
</tr>
<tr>
<td>Exhibit 10</td>
<td>Example of a test unit with components labelled</td>
</tr>
<tr>
<td>Exhibit 11</td>
<td>A good item: Multiple-choice item with plausible distracters</td>
</tr>
<tr>
<td>Exhibit 12</td>
<td>A flawed item: Multiple-choice item with a stem that is not self-contained</td>
</tr>
<tr>
<td>Exhibit 13</td>
<td>A good item: Multiple-choice item with self-contained stem</td>
</tr>
<tr>
<td>Exhibit 14</td>
<td>A good item: A closed constructed-response item with a detailed marking guide</td>
</tr>
<tr>
<td>Exhibit 15</td>
<td>A good item: Well-constructed short open-response item with a comprehensive marking guide</td>
</tr>
<tr>
<td>Exhibit 16</td>
<td>Summary of the advantages and disadvantages of different response formats</td>
</tr>
<tr>
<td>Exhibit 17</td>
<td>TRAPD model</td>
</tr>
<tr>
<td>Exhibit 18</td>
<td>PISA 2012 translation model</td>
</tr>
<tr>
<td>Exhibit 19</td>
<td>NAS translation model for Indo-Aryan languages</td>
</tr>
<tr>
<td>Exhibit 20</td>
<td>Two tests showing common item linking</td>
</tr>
<tr>
<td>Exhibit 21</td>
<td>Cluster rotation design used to form test booklets for PISA 2006</td>
</tr>
<tr>
<td>Exhibit 22</td>
<td>Example test item showing unique item identifier and scoring options</td>
</tr>
<tr>
<td>Exhibit 23</td>
<td>Example of an open-ended question</td>
</tr>
<tr>
<td>Exhibit 24</td>
<td>Example of a categorical question</td>
</tr>
<tr>
<td>Exhibit 25</td>
<td>Example of a question with a Likert scale</td>
</tr>
<tr>
<td>Exhibit 26</td>
<td>Example of a question using numbered boxes</td>
</tr>
<tr>
<td>Exhibit 27</td>
<td>Coverage and exclusions</td>
</tr>
<tr>
<td>Exhibit 28</td>
<td>Visual example of the effects of clustering using a multi-stage design</td>
</tr>
<tr>
<td>Exhibit 29</td>
<td>Four hypothetical surveys highlighting non-response bias effects on survey results</td>
</tr>
<tr>
<td>Exhibit 30</td>
<td>Example of an operation plan of an assessment project through to data collection stage</td>
</tr>
<tr>
<td>Exhibit 31</td>
<td>Example of an operation plan for a field investigator</td>
</tr>
<tr>
<td>Exhibit 32</td>
<td>List of key administration documents to be produced in different stages of the survey</td>
</tr>
<tr>
<td>Exhibit 33</td>
<td>Data cleaning in multi-stage data collection</td>
</tr>
<tr>
<td>Exhibit 34</td>
<td>Item and student scales under Classical Test Theory</td>
</tr>
<tr>
<td>Exhibit 35</td>
<td>Relationship between the latent and observed variables under Item Response Theory</td>
</tr>
<tr>
<td>Exhibit 36</td>
<td>Probability of success</td>
</tr>
</tbody>
</table>
List of Exhibits

<table>
<thead>
<tr>
<th>Exhibit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>Item-person map</td>
</tr>
<tr>
<td>38</td>
<td>Item difficulty scatter plot</td>
</tr>
<tr>
<td>39</td>
<td>Maximum likelihood estimates</td>
</tr>
<tr>
<td>40</td>
<td>Marginal maximum likelihood estimates</td>
</tr>
<tr>
<td>41</td>
<td>Summary of individual and population estimates</td>
</tr>
<tr>
<td>42</td>
<td>Example of proficiency scale</td>
</tr>
<tr>
<td>43</td>
<td>Questions for consideration when establishing an analysis plan</td>
</tr>
<tr>
<td>44</td>
<td>An example of a scatter plot</td>
</tr>
<tr>
<td>45</td>
<td>How socio-economic background relates to student performance in science (PISA 2006)</td>
</tr>
<tr>
<td>46</td>
<td>Dissemination methods</td>
</tr>
<tr>
<td>47</td>
<td>Example product specification sheet</td>
</tr>
<tr>
<td>48</td>
<td>Report writing considerations</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSS</td>
<td>Clustered sample size</td>
</tr>
<tr>
<td>CTT</td>
<td>Classical Test Theory</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential item functioning</td>
</tr>
<tr>
<td>EAP</td>
<td>Expected a-posteriori estimate</td>
</tr>
<tr>
<td>ESCS</td>
<td>Economic, social and cultural status</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum likelihood estimate</td>
</tr>
<tr>
<td>MML</td>
<td>Marginal maximum likelihood</td>
</tr>
<tr>
<td>NAEP</td>
<td>National Assessment of Educational Progress</td>
</tr>
<tr>
<td>NAS</td>
<td>National Achievement Survey</td>
</tr>
<tr>
<td>NCERT</td>
<td>National Council of Educational Research and Training</td>
</tr>
<tr>
<td>OCR/OMR</td>
<td>Optical character recognition/optical mark recognition</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Co-operation and Development</td>
</tr>
<tr>
<td>PIRLS</td>
<td>Progress in International Reading Literacy Study</td>
</tr>
<tr>
<td>PISA</td>
<td>Programme for International Student Assessment</td>
</tr>
<tr>
<td>PL</td>
<td>Parameter Logistic</td>
</tr>
<tr>
<td>PV</td>
<td>Plausible value</td>
</tr>
<tr>
<td>RMSA</td>
<td>Rashtriya Madhyamik Shiksha Abhiyan</td>
</tr>
<tr>
<td>SES</td>
<td>Socio-economic status</td>
</tr>
<tr>
<td>SLAS</td>
<td>State Level Achievement Survey</td>
</tr>
<tr>
<td>TIMSS</td>
<td>Trends in International Mathematics and Science Study</td>
</tr>
<tr>
<td>WLE</td>
<td>Weighed maximum likelihood estimate</td>
</tr>
</tbody>
</table>
Foreword

This handbook is intended as a guide and reference to Indian institutions and their staff that are engaged in, or that have an interest in, carrying out large scale assessment surveys of learner achievement.

Such assessments may be national or state level in scale, or at other levels of coverage. When adapted to the case in question, the principles described within this handbook have wide and general application.

The handbook has been developed by the National Council of Educational Research and Training (NCERT) specifically for the Indian context, drawing on international best practice to combine with and enrich the substantial experience of conducting such surveys already existing in the country.

In the combination of these complementary international and national experiences, NCERT sees the current and future development of high quality, objective and statistically rigorous large scale assessments of learning.

In particular, the technical focus of this handbook draws on the innovations which the Educational Survey Division (ESD) of NCERT has been introducing with the adoption of modern assessment survey techniques based on Item Response Theory and the progressive pursuit of international best practices in conducting large scale assessments of learning. Additional technical information and examples have been drawn from other assessments in India and elsewhere.

The chapters of this handbook provide an introduction to thirteen key areas of a robust, large scale assessment system. In planning, managing and conducting such assessment surveys, careful attention to each of these areas is important for the success of the overall undertaking.

This handbook is intended to be of interest to planners, managers and policy makers, as well as to academic and technical staff involved in the detailed work of carrying out assessment surveys. The handbook provides a sound and up-to-date overview of the field, for each of these categories of user.

Some of the content of this handbook is by necessity highly technical, however, and will be of direct interest primarily to those engaged in the hands-on work of the respective areas of the assessment process.

To complement this handbook, six brief, and more narrowly focused Guides are planned, which will deal in more technical detail with selected key areas of the assessment process, corresponding to the relevant chapters of this handbook. Whilst the present handbook is designed to be applicable to any large scale assessment survey in India, the planned Guides will be focused primarily on the conduct of the National Achievement Survey (NAS).

It is our sincere hope that this handbook will be of interest and of use to the many parties in India who have a stake in knowing what learners in our schools know and what they can do. The purpose of achievement surveys is to find that out, so that efforts may be best focused to help learners achieve their maximum potential and to enable the nation to benefit as a result.

(Prof. H.K. Senapaty)
Director, NCERT
Acknowledgements

This handbook was planned, designed, developed and produced by the National Council of Educational Research and Training (NCERT) with technical support provided through the Rashtriya Madhyamik Shiksha Abhiyan Technical Cooperation Agency (RMSA-TCA), funded by the UK Department for International Development (DFID).

The handbook has been developed and produced with the involvement of an extensive team, including experts drawn from NCERT, state education bodies, and through the RMSA-TCA.

Within the TCA consortium, managed by Cambridge Education, the contribution of assessment expertise, writing and editing of the handbook has been provided by the Australian Council for Educational Research (ACER), and document design and formatting expertise provided by DHA Communications.

The information and examples contained within the handbook draw on the experience of the Educational Survey Division (ESD) of NCERT and its partners in developing, implementing and reporting India’s National Achievement Survey (NAS) on behalf of the Ministry of Human Resource Development (MHRD). Further examples are also drawn from other sources, within India and internationally, acknowledged in the text.

Thanks are due to the institutional and individual backing provided to this undertaking, from the Minister, Secretary and Joint Secretary School Education and Literacy, MHRD, the RMSA-TCA Steering Committee, DFID-India, and the home institutions of each of the contributors to the process of planning, developing and producing this handbook.

Additionally, sincere recognition is made of the work of the very many people whose efforts have contributed to the development, conduct and reporting of assessment surveys over the years, in India and elsewhere, on whose accumulated experience and expertise this present handbook is based.

A list of direct contributors to this handbook appears at the end of the document.


ACER. (in press). *SEA-PLM assessment framework*. UNESCO.


publications
Introduction
Learning assessments can provide valuable information on the education system with regard to issues of access, quality, efficiency and equity [1].

They can answer questions, such as: What is the quality of the education that students are receiving? Is there equitable distribution of education? What factors might contribute to student achievement? Is there any change in student achievement over time? How much growth in educational outcomes is there between class (grade) levels? The information gathered from learning assessments can then be used by stakeholders to make decisions about educational policies and practices, with the ultimate aim of improving learning outcomes for students.

If learning assessments are to answer some of these fundamental questions, it is essential that assessment programs are of high quality. To be useful in the policy process, an assessment program should be ongoing and long-term with a clearly defined strategic focus. Of course, if assessment findings are to provide valid and reliable information to stakeholders, the assessment must be technically robust. The robustness of an assessment program is one of the important factors in determining whether or not assessment data are used in policymaking [2].

Developing a high-quality learning assessment program is a technically challenging and resource-intensive activity. This handbook is intended to support the process of developing and implementing large-scale assessments in India. It provides information on core areas to consider when developing learning assessment programs at the national or state levels.
1.1 The assessment context in India

Over the past decade, India has made great progress towards increasing access to education [eg 3]. More recently, there has been a greater focus on also ensuring the quality and equity of education in India. If India’s education system is to improve the learning outcomes of students, it is essential to understand what and how learning occurs, and what improvements can be made. Learning assessments can play an important role in understanding and improving the quality of education, by providing information on students’ learning outcomes and on the background and environmental factors that may influence these outcomes.

In order to promote access, equity and quality in elementary and secondary education, the Government of India implemented the Sarva Shiksha Abhiyan and Rashtriya Madhyamik Shiksha Abhiyan (RMSA) programs. As part of these programs, the National Achievement Survey (NAS) was implemented in 2000 by the Ministry of Human Resource Development and the National Council of Educational Research and Training (NCERT). The NAS provides information about Class III, V, VIII and X student achievement in various subjects including language, mathematics and science. It also provides information on changes in achievement across classes and on trends over time.

More recently, State Level Achievement Surveys (SLAS) have been implemented to complement the NAS by providing finer detail about student outcomes at the state and district levels. This handbook has been written in order to support large-scale learning assessment programs in India, including programs such as the NAS and SLAS.
This handbook is comprised of 13 chapters (in addition to this chapter), each addressing a core aspect to consider when developing a robust learning assessment. The handbook is intended to support educational policymakers and planners in India, both at the state and national levels, evaluation and assessment experts, data analysts and psychometricians, test developers and other personnel involved in planning and conducting large-scale assessment projects in India.

Readers of this handbook are likely to find certain chapters more relevant to their roles than others. While some chapters will be more applicable to policymakers, and others more relevant to psychometricians or test developers, it is recommended that where possible, all team members become familiar with the different components of an assessment program. Each of the 13 areas discussed are significant in developing a robust assessment. These 13 interconnected areas can be seen in Exhibit 1. The first of the 13 topics in this handbook is on policy goals. Chapter 2 discusses some of the different purposes and features of large-scale learning assessments, and in particular, the policy goals and issues that learning assessments may address.

Essential to the success of any large-scale assessment is a dedicated project team with clear roles and the expertise required. Chapter 3 discusses the project team and the infrastructure requirements for developing a high-quality assessment program.

Part of the quality assurance process for large-scale assessments is to establish a set of technical standards that, if met, will establish the technical credentials of the project. Chapter 4 presents a set of technical standards that are applicable within the context of large-scale assessments in India and provides examples of best practice.

A cornerstone of a robust assessment program is the assessment framework. The assessment framework lays out the principles upon which an assessment is built. Chapter 5 discusses why and how frameworks are developed and the different elements of an assessment framework.

The assessment framework forms the basis for the development of the assessment items (questions). Chapter 6, on item writing, provides information on what characterises good assessment items and test instruments, and how this quality is achieved through the item development process.

Chapter 7, on equivalence in multilingual assessments, discusses translation and linguistic quality-control principles and models. Ensuring that tests delivered in more than one language are equivalent is a particularly relevant issue for India, where tests are often administered in multiple languages.

Once assessment items have been developed and translated, these are then used to compile the tests. Test design is discussed in the first part of Chapter 8. The second part of Chapter 8 describes another fundamental component of instrument construction, the questionnaire design. Questionnaires are often an important component of large-scale assessment programs, providing complimentary data to achievement tests.
If the assessment design is a sample survey, then scientific sampling techniques must be utilised to enable valid inferences to be made about a population of students. Chapter 9 includes information on sampling principles and techniques in order to ensure the appropriate level of statistical precision and validity.

The production of assessment materials and the field operations are outlined in Chapter 10. This chapter discusses the activities that need to occur before, during and after the testing day. It includes information on how standardised test administration procedures can be implemented so that each student is assessed under the same conditions.

Once the assessment data have been collected, the data need to be entered, cleaned, verified and validated. Chapter 11 discusses the best practices for these different stages of data management. Assessment results are better interpreted, and therefore more useful, if they are reported on a scale. This requires the use of advanced statistical techniques. Chapter 12 discusses scaling methodologies, an important consideration for proper interpretation of data.

Chapter 13 explores the statistical data analysis methods that can be used to help ensure the results are accurate and reliable. It discusses the types of analyses that can be conducted so that assessment results can inform policy and practice.

Once the data have been analysed, they need to be reported on and communicated to stakeholders. Chapter 14 discusses some of the considerations and options for the reporting and dissemination of assessment information, so that assessment results are understandable and useful to stakeholders. As Chapter 14 outlines, it is important that dissemination approaches are first considered at the outset of an assessment program.

Throughout the handbook, examples are provided from assessment programs in India and from other large-scale national and international assessment programs. Where tables and figures are included in the handbook, these are all referred to as ‘exhibits’. A glossary of technical terms, as well as references, is included at the end of each of the chapters.

Each chapter concludes with a checklist, which readers can use to review their own assessment programs. Readers can reflect on whether they have considered and implemented some of the key points discussed in the chapter. While these cover some of the important aspects of a learning assessment, these should not be taken as comprehensive instructions. As previously mentioned, developing and implementing large-scale assessments is a complex process, and all aspects cannot be discussed in one handbook. This handbook is just one of many resources that should be utilised and consulted when making decisions about large-scale assessments.
Exhibit 1: 13 key areas of a robust large-scale assessment program

A robust large-scale assessment program

Policy goals and issues
Project team and infrastructure
Technical standards
Assessment framework
Item writing
Equivalence in multilingual assessments
Test and questionnaire design
Sample design
Materials production and field operations
Data management
Scaling methodology
Analysis for reporting
Reporting and dissemination

Source: Adapted from ACER [4]
<table>
<thead>
<tr>
<th><strong>Glossary</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Psychometrician</strong></td>
</tr>
<tr>
<td><strong>Test developers</strong></td>
</tr>
</tbody>
</table>
References


Policy goals and issues
Policy goals and issues

Large-scale learning assessments are usually the result of some sort of policy initiative: they are initiated and funded by a policymaking authority such as a ministry of education.

Policymakers, therefore, expect a return from the assessment program that they fund. Designing and implementing a large-scale assessment requires a big investment in terms of time and resources. Articulating the purpose of the assessment and identifying how the assessment may inform policy discussions and educational practice are absolutely fundamental to the assessment design. There is no point in collecting data unless the data can usefully inform the policy process and/or have some formative impact on educational practice.

This chapter will firstly discuss the initial major considerations when thinking about an assessment program: What are the current national or state educational priorities that the assessment will inform? What is the purpose of the assessment? Who are the major stakeholders? It will then go on to discuss considerations around the scope of the assessment: Who will take the assessment? In what subjects? How often? Following this, some further elements of assessment design are outlined, including measuring background characteristics that might have an effect on educational outcomes, and logistical considerations about how the assessment is to be implemented.
2.1 Education priorities

An assessment needs to address educational policy issues. Therefore, determining educational priorities is a first step in the development of a large-scale educational assessment. National or state education priorities are often explicit in legislation and regulations but also often implicit in various policy documents (e.g., white papers, program documents, and framework documents) and public statements made by policymakers.

The Right of Children to Free and Compulsory Education Act 2009 [1], for example, enshrines the right of access to education for all Indians aged between 6 and 14. Article 8 of the Act [1] states that the appropriate government must ‘ensure good quality elementary education’. This access and quality of education is to be extended to all children. As outlined in Article 2 [1], this right includes children ‘belonging to the Scheduled Caste, the Scheduled Tribe, the socially and educationally backward class or such other group having disadvantage owing to social, cultural, economical, geographical, linguistic, gender or such other factor’. So, along with the fundamental right to education, there is a notion that the education provided should be purposeful and of a quality to deliver the desired outcomes. From this one Act it can be seen that the educational policy priorities of India include access to education, quality of education, and equitable distribution of educational outcomes.

Educational priorities are not always public and straightforward. For various reasons, policymakers may not always want to draw public attention to particular educational issues. Issues around the quality of private versus public management of schools, or the mathematical literacy of primary school teachers, for example, are often sensitive issues. Often it is only behind closed doors that the hot topics of educational policy debates can be identified. It is the duty of both the policymakers and the educational assessment and evaluation experts to clearly determine the educational priorities of the day, so that these priorities can be foremost in mind in designing large-scale assessments.
2.2 The purpose of assessment

While large-scale assessments can serve different purposes, assessments are generally most effective when they are tailored to specific purposes. Designing a large-scale assessment first requires that there is a sound understanding of the purpose. Here are some examples of different purposes for assessment:

- An examination can be used to rank students in order of ability, so as to select the most able students for the next stages of education.
- A census-based monitoring study can be used to provide feedback to students, parents and teachers on how each student compares to others in the same class.
- A sample survey assessment can be used to efficiently quantify differences in learning between populations and sub-populations, such as differences between boys and girls; between urban and rural dwellers; between social groups; between those attending private schools and those attending government schools; or between those students schooled in English and those schooled in Hindi medium of instruction.
- A standardised diagnostic assessment administered to students can help teachers place each student in a class where the level of teaching is appropriate to each student’s current ability.
- A sample survey assessment can be used to track learning outcomes over time for a particular class. For example, the National Achievement Survey run by the National Council of Educational Research and Training (NCERT) currently do this for Classes III, V, VIII and X.
- A vertically linked assessment can be used to determine how much growth in learning is taking place between two learning levels, for example between Class III and Class V.

The purpose of an assessment affects many aspects of the study design. Exhibit 2 provides examples of two assessments, each with a different purpose and illustrates how this affects aspects of the study design.

1 Information on vertical linking can be found in Section 12.2 of this handbook.
### Exhibit 2: Assessment purposes and study design

<table>
<thead>
<tr>
<th>Assessment purpose and study design</th>
<th>Example assessment 1</th>
<th>Example assessment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td>Ranking students to select the most able students for the next stages of education</td>
<td>Quantify differences in learning between sub-populations</td>
</tr>
<tr>
<td><strong>Assessment type</strong></td>
<td>Examination</td>
<td>Sample-based assessment</td>
</tr>
<tr>
<td><strong>Importance to test-takers</strong></td>
<td>High-stakes; high risk of irregularities or dishonesty in test situation</td>
<td>Low-stakes; low risk of irregularities or dishonesty in test situation</td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td>Highly secure</td>
<td>Low security</td>
</tr>
<tr>
<td><strong>Scope</strong></td>
<td>Coverage of topics limited to the content of a single test form</td>
<td>Wider coverage of topics possible due to the possibility of administering multiple test forms</td>
</tr>
<tr>
<td><strong>Consistency</strong></td>
<td>Needs to be fair to the individual, so it is necessary that everyone gets exactly the same test</td>
<td>Needs to make accurate inferences to sub-populations, so it is not necessary that everyone gets exactly the same test</td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td>Simple, percent-correct scores are sufficient</td>
<td>Complex statistical methods are required</td>
</tr>
<tr>
<td><strong>Reporting</strong></td>
<td>Reports at the individual level</td>
<td>Reports at the level of populations and sub-populations</td>
</tr>
</tbody>
</table>
2.3 Major stakeholders

The major stakeholders should guide the overall design of the large-scale assessment. Typically, major stakeholders are identified by the funders or sponsors of the assessment and this is usually some form of government ministry or department of education. Major stakeholders may include:

- government representatives at a senior level, eg minister, deputy minister, secretary, joint-secretary, undersecretary of education
- funding or development partners, such as international banks, donors or contributing non-governmental organisations
- educational planning advisors
- education policy developers
- curriculum developers
- teacher educators
- school monitoring officials.

Other stakeholders may include members of teacher unions, academics, experts in curriculum or pedagogy, economists and educational researchers. If the large-scale assessment is designed to provide individual results, then teachers, students and parents are major stakeholders. Depending on the nature of the assessment, district officials and school heads may be considered stakeholders as well.

Having identified the major stakeholders, it should be considered how their interests will be represented. One way to do this is to establish representation in a project steering group.
2.4 Project steering group

Stakeholders can help to clarify the purpose of the assessment and their reporting requirements from the assessment. One way to gather this information from stakeholders is through a project steering group. A steering group is a committee that meets to discuss the overall direction and priorities of the assessment program.

The steering group might meet just two or three times at the beginning of the project, then once every four to six months for progress reports and to discuss issues. They may meet another two or three times to discuss the results and to help plan the communication strategy.

The steering group should be made up of people appointed to represent stakeholders as well as people representing the project’s overall management and technical management direction. This representation could include the following roles:

- project manager/coordinator
- assessment development manager/coordinator
- logistics manager (especially at the beginning of the project)
- sampling, scaling and data manager
- communications manager.

All of these roles could usefully contribute to design discussions and highlight any technical implications of suggestions from stakeholders.

The project steering group should:

- define the purpose and scope of the assessment
- formulate or understand the research questions that the assessment is designed to answer
- anticipate the impact of the assessment results on educational policy and practice
- have a sound understanding of the budget and human resources available to undertake the assessment program
- specify the assessment design elements.

There are often other advisory groups for assessment programs other than the steering group. There may be technical expert groups to address methodological issues, subject expert groups to address content issues, and monitoring committees.

---

2 These examples refer to the roles outlined in Chapter 3.
2.5 Scope of the assessment

This section discusses the major aspects determining the assessment scope. More information is provided in other chapters, particularly on sampling (regarding populations) (Chapter 9), assessment frameworks (Chapter 5) and test design (Chapter 8).

2.5.1 Target population

All assessments have a target population. Class X Board examinations, for example, are targeted to the population who have just completed Class X. The function of sample-based assessments is to make inferences about outcomes to populations. The target population can be defined in various ways but for educational assessments these are generally limited to two: age-based populations and class- or grade-based populations. The advantages of class-based populations are:

- The students are clustered together (in schools, classes and sections) so the sample frame is easy to construct and the data are easy to gather.
- The results have a clear frame of reference (for example, everyone knows what is meant by ‘students in Class X’ for example, but ‘15-year-olds’ might be in Class IX, X or XI or may not even be attending school).

When defining the target population it should also be considered whether two or more populations are to be compared concurrently or over time. When monitoring the outcomes from an educational system, a lot of value is added when the program monitors the educational growth between classes. For example, the growth in reading comprehension between Class V and Class VIII or the growth in financial literacy between Class VIII and Class X can be measured.

More detail on defining the target population is provided in Section 9.2 of this handbook.

2.5.2 Sub-populations of interest

As was outlined in Section 2.1, equitable outcomes for sub-populations defined by gender and social category are key policy considerations in India. To monitor outcomes for these sub-populations, reliable data on students’ gender and social category should be gathered. Examples of sub-populations of interest include those defined by:

- gender
- social category
- socio-economic status (SES)
- language of school instruction
- attendance at school
- attendance at a government-aided or private school
- attendance at a rural or urban school
- particular disabilities.
2.5.3 Sample coverage

The coverage of the sample is another topic for the project steering group to discuss. For example, if the steering group has a research question about the outcomes of Class V students, they will need to consider whether this means all Class V students or particular groups of Class V students. For example, the sample may or may not cover students:

- in private schools
- in schools for students with special needs
- housed in special facilities such as prisons or hospitals
- learning by distance education
- in very small remote schools
- in religious minority schools
- who are homeschooled.

2.5.4 Frequency of assessment

Large-scale assessments are most useful when they are repeated over time so that the results can be compared and outcomes monitored. The frequency of assessments will also depend on what the purpose of the assessment is.

Take, for example, a census-based diagnostic assessment such as the National Assessment Program – Literacy and Numeracy. This is undertaken in Australia every academic year for all students in Years 3, 5, 7 and 9 of schooling. Because the purpose of the assessment is diagnostic, it is important to provide regular information to students, their parents and teachers about the individual’s basic educational strengths and weaknesses.

Large-scale sample surveys, such as the National Achievement Survey (NAS) in India, are aimed at measuring outcomes at the population and sub-population levels. Therefore, they are often cyclic in nature: students in different class levels are assessed every three to four years.

The benefits of having frequent assessments to gather regular information must be balanced against several considerations, including:

- the cost of gathering the information
- the capacity of the assessment project team to design, implement, analyse and report with a regular turnaround
- the fact that in many educational systems, even when large-scale policy and practical initiatives are implemented, any positive changes often take a long time to emerge.

It is important that the frequency of assessments allows sufficient time for all aspects of an assessment program (as shown in Exhibit 1) to be implemented to a high standard.
2.5.5 Target users of assessment results

Again, the purpose of the assessment is essential to determining who will use the assessment results. Diagnostic assessment results will frequently be used by teachers but also by the parents and often the students themselves. The results of population surveys are more frequently of interest to policymakers, researchers, academics, curriculum planners and teacher trainers. Chapter 14 discusses the need to consider the target audiences when making decisions about how to disseminate the assessment results.

2.5.6 Domains to be assessed

A domain here refers to an area of inquiry where the test-taker is required to demonstrate cognitive skills and/or knowledge. Initial considerations by major stakeholders should include discussions on the domains or subjects to be assessed. Is the assessment to be linked to a curriculum subject such as mathematics or science? Or is the assessment cross-curricular or non-curricular, like ‘problem-solving’ or ‘financial literacy’. These considerations are discussed in more detail in Chapter 5.
2.6 Design elements of the assessment

This section briefly introduces some of the main design elements of an assessment. More information is provided in other chapters on assessment frameworks, item writing, test design, sampling and field operations.

2.6.1 Mode of assessment

The mode of assessment will be determined by the purpose of the assessment, the nature of the domain to be assessed and logistical considerations. Multiple-choice tests are very efficient at gathering certain types of data but it is very difficult to measure higher order thinking skills and impossible to measure skills like expressive dance with multiple-choice questions.

Assessments requiring the test-taker to write essays can be very good at assessing writing skills but they take a long time to mark and a lot of effort to train and monitor raters for reliability.

One-to-one interview-based assessments can be very good to elicit information about the skills of younger children. However, these also require a lot of time and training which can be a prohibitive expense for a large-scale assessment.

Assessments can be mixed-mode, containing both multiple-choice and essay-type items, for example. Chapter 6 on item writing provides more detail on item response types.
2.6.2 Contextual considerations

While comparative information on educational outcomes is important, information on contextual factors that would be expected to influence these outcomes is significantly more useful for policymakers. These contextual factors, often referred to as background variables, can be organised around several levels:

- the individual context of the student
- the home environment of the student
- the school and classroom level
- the wider community.

Some of the most important contextual factors that have been shown to influence educational outcomes include:

- the socio-economic status (SES) of parents
- the educational resources within the home
- the degree of ongoing systematic professional development of teachers in the school.

Section 5.4 of this handbook outlines an approach to developing a contextual framework for an assessment. Section 5.4 and Section 8.2 discuss choosing among the many contextual factors that may be of interest to policymakers to measure.

2.6.3 Logistical considerations

Major stakeholders will need to be conscious of the logistical implications of a large-scale assessment. As noted earlier, the person managing logistics should be present at the project steering group meetings, especially at the beginning of the project when some of the substantive decisions are to be made.

There are many important logistical considerations. These include considering the following:

- Who will gather the data?
- How will the measures be recorded?
- What is the required turnaround time between data collection and reporting?
- Does the project team have enough staff to undertake data entry?
- Does the project team have the technical skills and time to undertake the assessment tasks within the required timeframe?
- Is outsourcing of some functions a possibility?
- Are there any technological solutions available to assist – such as scanning of data – that might work well to manage data and reduce errors? And what are the relative costs of each option?

Careful planning is required to ensure that the assessment program is designed to be effective, robust, coherent and achievable. It is important that this planning takes into account the budget, time and human and physical resources available (as is discussed further in Chapter 3).
2.7 Considerations for policy implications and issues

This chapter has largely focused on issues that arise when first considering assessment design and implementation alternatives. However, engagement with the policy process and major stakeholders is critical throughout the life cycle of the assessment program.

Shifting political and policy-related priorities can cause the scope of the assessment to widen or shift. Stakeholders may suddenly develop a strong interest in particular educational groups – such as private schools, the homeschooled, or students with special needs – and ask for them to be represented in the program.

Importantly and inevitably, the members of the project steering group will change over time. Different people will bring different perspectives and educational foci. The project steering group must meet regularly to ensure continuity in direction. It is important that stakeholders provide constructive feedback to the assessment team. It is also important that stakeholders are informed of any unexpected developments or delays.

When the analysis of the assessment results is complete or at least underway, the likely impact of the results on educational policy and practice should be critically and deeply discussed among the major stakeholders. Yet in 2013, a systematic review of literature aimed at determining the impact of large-scale assessments on the education policy process revealed that little material could be found about the impact of large-scale assessments that have been undertaken in India [2].

There are, though, many examples from throughout the world of large-scale assessments having an impact on the education policy process. The following examples in this section are taken from a study that examines the use of large-scale assessment results in education policy in the Asia-Pacific region, including low, middle and high-income countries [3].

Most frequently, assessment results are used in policymaking at the monitoring and evaluation stage to legislate further monitoring and evaluation activities. This often included the establishment of assessment bodies and the legislation of regular conduct of assessments. For example, Vietnam used national assessment results to evaluate policies that had been implemented prior to the assessment, including policies related to curricular reform, class size and teacher contact hours [4].

Assessment results are also used by policymakers at the policy implementation stage of the policy cycle, when assessment data are used to improve the ways in which an initiative is implemented or targeted on the ground. For example, student assessment results in Madya Pradesh helped to support the implementation of the Learn to Read initiative. The data allowed the provision of teacher coaches and other supports to be effectively targeted to districts, schools and teachers [5].

Assessments are most frequently used to inform system-level policies, which provide a framework for evaluation systems and operations. For example, Iran’s participation in the Trends in International Mathematics and Science Study (TIMSS) led to the use of the TIMSS curricular framework in the development of test items for the country’s own assessment uses [6].
The establishment and reform of curricular and performance standards aims to provide a common framework and context for the interpretation of assessment results. For example, Kyrgyzstan’s lower than expected results in the Programme for International Student Assessment (PISA) in 2006 helped to justify the government’s curricular reforms to emphasise ‘modern skills and competencies’ in line with those assessed by PISA, and an expectation of higher student performance [7].

Assessments can also have an impact on policy formulation, though this is less frequent than use at other stages of the policy cycle. Impact on policy formation occurs through the use of assessment results to inform the design and formulation of policy options and strategies. For example, policymakers in Japan legislated a daily morning reading session as a response to declining PISA results in literacy [8].

There is less available evidence of large-scale assessments having an impact on teaching and learning policies, which are aimed at specific school- and classroom-level practices. However, an examination of the available evidence shows that when assessments results inform teaching and learning policies, these policies most frequently focus on in-class and enhanced learning strategies [3].
Large-scale assessments can provide valuable empirical evidence to inform policy discussions around educational priorities. This chapter has pointed to the importance of determining the current educational policy priorities at the very outset of considering an assessment.

The purpose of the assessment and research questions that guide the development of the assessment should be clearly articulated with a view to designing an assessment program that is of maximum value to policymakers and practitioners. To ensure this happens, it is important to identify the major stakeholders and include them in deliberations on the purpose of the assessment and its design.

The creation of a project steering group is a useful way to involve stakeholders in assessment decisions. The steering group would guide the scope of the assessment, determining what and who is being assessed, and how often. They would be involved in details of the assessment design, influence the choice of background factors to measure and make overarching logistical choices regarding how the program is to be run. They would also ensure that the methods and logistics were aligned with the budget and that a human and physical resource strategy was in place to implement the program.
## Checklist: Policy goals and issues

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Project steering group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have the major stakeholders been identified?</td>
</tr>
<tr>
<td>2. Has a project steering group been established?</td>
</tr>
<tr>
<td>3. Does the steering group have representation from all major stakeholders?</td>
</tr>
<tr>
<td>4. Is the project manager or coordinator a member of the steering group?</td>
</tr>
<tr>
<td>5. Is there sufficient representation of the project’s technical leaders on the steering group?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessment purpose and scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. Has the purpose of the assessment been articulated and documented?</td>
</tr>
<tr>
<td>7. Have the research questions been formulated and documented?</td>
</tr>
<tr>
<td>8. Has the connection between the research questions and educational policy priorities been clearly documented?</td>
</tr>
<tr>
<td>9. Have the populations and sub-populations of interest been identified and documented?</td>
</tr>
<tr>
<td>10. Have the domains (subjects) to be assessed been determined?</td>
</tr>
<tr>
<td>11. Have the contextual factors of interest been determined?</td>
</tr>
<tr>
<td>12. Have the key groups and individuals that will use the assessment results been identified?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design and implementation issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Has the sample coverage been determined and documented?</td>
</tr>
<tr>
<td>14. If the assessment program is to be recurrent, has the frequency of assessment been determined?</td>
</tr>
<tr>
<td>15. Have the broad logistics of assessment development, data collection and data processing been determined?</td>
</tr>
<tr>
<td>16. Are the human resources in place, or have the human resources been planned?</td>
</tr>
<tr>
<td>17. Is the budget sufficient, or is a sufficient budget foreseen?</td>
</tr>
<tr>
<td>Term</td>
</tr>
<tr>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Census-based assessment</td>
</tr>
<tr>
<td>Contextual factors</td>
</tr>
<tr>
<td>Diagnostic assessment</td>
</tr>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>High-stakes assessment / low-stakes assessment</td>
</tr>
<tr>
<td>Monitoring study</td>
</tr>
<tr>
<td>Sample frame</td>
</tr>
<tr>
<td>Standardised assessment</td>
</tr>
<tr>
<td>Sub-population</td>
</tr>
<tr>
<td>Target population</td>
</tr>
<tr>
<td>Vertically linked assessment</td>
</tr>
</tbody>
</table>
References

1. Right of Children to Free and Compulsory Education Act 2009 (India).


Project team and infrastructure
An important consideration at the outset of designing an assessment program is organisation of the resources that are needed to design and implement the program.

It is essential that human and physical resources are planned, that budgets are calculated and that strategies for engaging managers and other staff are developed. The topics discussed in this chapter are highly situational.

### 3.1 Assessment centre

For efficient and effective design, management and implementation of a large-scale assessment, it is useful to establish some sort of assessment centre. The assessment centre may be part of the National Council of Educational Research and Training (NCERT), or it may be a project team within a research agency or university, for example. For state assessments, the assessment centre may be located in the State Council of Educational Research and Training, State Institute of Education, Institute of Advanced Studies of Education or a College for Teacher Education, according to the management set-up and decisions of the state or union territory in question.

The way that the assessment centre is organised will depend upon the institution in which it will sit, the funding source for the assessment, and any organisational or bureaucratic constraints present. One very important consideration, however, is that large-scale assessments invariably involve an enormous amount of organisation of processes, staff and logistics. There are also often very demanding timeframes involved, with many finish-to-start task dependencies. For example, sampling cannot begin until sample frame information is gathered; assessment booklets cannot be printed until the translation is finished. It is highly recommended that assessment organisers establish a core team within an assessment centre to manage the assessment project or program on a full-time basis, and with the minimum possibility of being sidetracked by other work.
3.2 Budget and source of funding

While the program budget and its source are very important to the design and implementation of any large-scale assessment, only limited attention can be devoted to these issues in a handbook such as this. This is simply because the funding structures, amount of funding available and the costs of different large-scale assessment choices vary enormously across India.

The financing of a large-scale assessment is often provided by some sort of governmental budget line. Estimating the costs is something that would have to be done by the senior members of an assessment team, once all the design parameters had been decided. The remainder of this chapter will focus on some of the fundamental staff, physical resource and service provision costs that should be considered.
3.3 Team organisation

There is considerable variety in how large-scale assessment teams are organised. Often, an organisation considering implementing a large-scale assessment will have existing structures which will serve to accommodate the new program. This section is therefore not prescriptive in nature. The organisational chart shown in Exhibit 3 is simply an exemplar. It is based on the core management roles presented in Section 3.4. This exemplar would be one way of organising the assessment team if an assessment centre was initially being designed.

In the organisational chart, the core management roles are presented in solid boxes. The responsibilities of these managers and the ideal experience of these people are outlined in Section 3.4. The staff, contractors or organisations that each manager would be responsible for are listed below the core management roles. Primary responsibilities are indicated with solid arrows. Dashed arrows represent secondary interactions. For example, while the sampling, scaling analysis and data manager will primarily be responsible for the data analysis, the assessment development manager and the communications manager will require significant interaction with the data analysts to ensure that reporting on the assessment is carried out as planned and to ensure that what is communicated is technically accurate.

Exhibit 3: Example of an assessment team organisational chart
3.4 Core team management

Key management roles are outlined in this section. An overall project manager is essential to any assessment team. Depending on the type and scale of the assessment, consideration should be given to designating functional management roles for:

- assessment development
- translation and linguistic quality assurance
- logistics
- sampling, scaling, analysis and data management
- communications.

These five management roles are outlined below. All of these aspects of an assessment program require different types of training and expertise. Determining whether some or any of these management roles can be shared is a matter of considering the scale of the program and the continuity of the various workloads. It is important to consider whether there is enough work over the period to keep a manager employed full-time in one or more of these areas.

3.4.1 Project manager

The project manager, often referred to as project coordinator, takes responsibility for the implementation of the project, working in liaison with relevant national or state educational institutions.

The position of project manager will usually be full-time, but could be part-time depending on the availability of other staff such as an administrative assistant and a data manager. It is strongly recommended that the project manager works on the project on at least a half-time basis.

Ideally, the person who is appointed as project manager has:

- previous experience in planning, organising and conducting large-scale surveys
- skill in managing a team of project staff who carry out multiple tasks often needing simultaneous attention
- a high level of oral and written communication skills
- an understanding of technical English (as many technical resources will be in English)
- familiarity with sampling, datafile structures, data management and data processing procedures
- sufficient knowledge and confidence to represent the project at steering group meetings where aspects of the project will be discussed
- knowledge of and the confidence to deal with government agencies, school heads, parents and teachers within India.
It is desirable that the person also has previous work experience in educational assessment and has a strong understanding of the statistical and measurement theories as introduced in this handbook. The project managers, along with the other team managers, will develop the technical standards of the program and ensure ways to meet them (see Chapter 4 for information on technical standards).

### 3.4.2 Assessment development manager

The assessment development manager will focus on the content of the assessment and any background questionnaires that are developed. This manager will be responsible for:

- the development of the assessment framework and the contextual framework
- training the assessment test developers and contextual questionnaire developers
- developing quality assurance processes
- constructing the assessment codebooks
- determining the best overall test design.

Chapters 5, 6 and 8 of this handbook deal with these topics.

### 3.4.3 Translation and linguistic quality assurance manager

If the assessment tools are developed in one language and then delivered in one or more different languages, a translation and linguistic quality assurance manager will be required. The risks to an assessment when there is poor quality translation should not be underestimated. When the measurement properties of an item change during the process of translation – for example, an item becomes harder because the translator has used a more advanced vocabulary – it is no longer useful to make comparisons with the original.

The translation and linguistic quality assurance manager will ideally have a formal background in translation. This manager would be responsible for:

- developing the linguistic guidelines for test developers
- reviewing items for translatability issues
- developing the translation guidelines
- ensuring the documentation of translation issues
- designing and implementing a linguistic quality control process
- managing the translation team(s)
- undertaking linguistic review of items where statistical concerns are raised.

Chapter 7 provides detail on translation and linguistic quality assurance processes.
3.4.4 Logistics manager

The logistics of a large-scale assessment are varied, many and often complex. A manager can be occupied full-time throughout the assessment process just looking after logistics. Chapter 10, on materials production and field operations, introduces the topics around logistics.

The tasks for the logistics manager would include:

- managing the materials production (layout, graphics, desktop publishing)
- sourcing the printing
- packaging distribution and return of assessment materials
- developing field operations
- training the field investigators
- organising quality assurance visits to assessment sites
- scheduling assessment sessions
- gathering feedback on the quality of field operations and taking steps to improve processes where necessary.

The logistics manager may also be involved in arranging data entry and/or scanning of booklets or optical character recognition/optical mark recognition (OCR/OMR) sheets.

3.4.5 Sampling, scaling, analysis and data manager

The complex technical aspects of a large-scale assessment require a highly skilled technical expert to manage tasks around sampling, scaling and analysis (see Chapters 9, 12 and 13 for information on these topics). Such a person will generally require a background in psychometrics and statistics for social science. This person would probably also be responsible for general data management as this requires a similar skill set (see Chapter 11 for information on data management).

Tasks that would be the responsibility of this manager (or the staff reporting to this manager) would include:

- obtaining accurate sample frame data
- designing the sample
- drawing and documenting the sample
- cleaning and ensuring consistency in the results data
- analysing questionnaire data from field trials and recommending main survey item selection, including the validation of questionnaire scales
- scaling the data from the main survey, possibly using Item Response Theory models
- applying sample weights
- constructing replicate weights
- constructing the assessment database
- performing complex statistical analyses.

The data manager may also be involved in arranging data entry and/or scanning of booklets or answer sheets.
3.4.6 Communications manager

The final area where there is enough work to consider engaging a senior manager is in communications. This person may have a background in communications, editing or media. They must have strong communication skills and be able to develop and manage strategic plans. Tasks for the communications manager may include:

- the design of a communications strategy
- management of report writing
- editing
- development of policy briefs
- seeking and facilitating media opportunities
- providing appropriate feedback to stakeholders.

Communications should be planned and implemented at all stages of the assessment program. During the initial discussions around the purpose and scope of the assessment, a skilled communications specialist will be thinking ahead about how to effectively communicate with stakeholders to inform them about the assessment program and its benefits to them.

The communications manager also works with other teams to help them inform key actors in the assessment process itself. For example, the production of training videos for field investigators, the establishment of a live chat helpdesk for school heads, or the production of participation certificates and results reports for students.

Planning and managing communications is usually at its apex when preparing for assessment results to be released. Especially at this stage, the communications manager may require the help of a small team for report writing, graphic design and production. Chapter 14 of this handbook outlines the main processes and choices to be made at the reporting and dissemination stage.
3.5 Capacity development

It is inevitable that that assessment team will be required to develop new skills and expertise to deliver a high-quality assessment program. Even very experienced and technically expert assessment teams need to continually develop new skills because of innovations such as:

- Development in statistical and measurement theories
- The introduction of new technologies and tools (for example, scanning and recognition of students’ free-hand written responses; use of the Unified District Information System for Education or the All India School Education Survey for sampling within India)
- The introduction of new assessment technologies (for example, computer-based assessments; collection of background information using mobile technologies).

At the beginning of the assessment program and at regular times throughout, a useful approach is to undertake i) an existing capacity analysis, and ii) a capacity needs analysis. The degree to which there is a mismatch between these two analyses indicates the extent of the capacity development needs. This approach is commonly known as a gap analysis, because it focuses on the end goal as a starting point and determines what needs to be done to fill the gap.

All capacity development involves learning activities. Determining appropriate learning activities is fundamental to any capacity development plan as not all such activities will be available, cost-effective and timely. Learning activities may include:

- Internships
- Study visits
- Short courses
- Long courses (for example, postgraduate diploma)
- Customised technical training provided by consultants.

Where learning opportunities are unavailable or impractical, outsourcing should be considered (see Section 3.6).

Capacity development is not limited to core assessment team members. There will be many members of the supporting staff who may require training in the specifics of the project, including:

- Test developers
- Translators
- Field investigators
- Data management staff.

Training of such staff will usually be undertaken by core staff. Where large numbers of people require training (for example, field investigators) then the use of train-the-trainer models, detailed manuals, webinars and video training methods might be utilised.
3.6 Outsourcing

Typically, a large-scale assessment will require some outsourcing of roles. This may be because the expertise required for specialised tasks is not available internally or through networks. The assessment centre may need support, for example, with item writing, translation, linguistic quality assurance, graphic design, sampling, psychometric analysis, editing or publishing.

Outsourcing may also be necessary because of the scale of the work involved. For example, the assessment centre may need to procure printing, distribution, field invigilation, marking, or scanning of assessment materials.

Advantages of outsourcing include:

• access to skills and facilities not available in the assessment centre
• increased capacity to process work at peak times
• fixed price and timeframe for delivery.

The costs and risks of outsourcing include:

• time needed to specify criteria, identify contractors, obtain quotes and negotiate price
• management of contractor delivery may become problematic
• outgoing budget costs.

3.6.1 Insourcing versus outsourcing

For the purpose of this handbook, insourcing refers to the type of extra staff support that is necessary to buy-in, usually on a short-term basis. It is similar to outsourcing in that it is providing auxiliary staff and skills. The difference between outsourcing and insourcing, however, is that with outsourcing, generally an expert individual or team is contracted and those experts undertake the professional responsibility of completing the job. With insourcing, generally staff are brought in, trained, and managed by core team members.

The difference between outsourcing and insourcing can be helpful when considering any organisational constraints around contract management on the one hand and hiring temporary staff on the other hand. As an example, data entry might be entirely outsourced and handled by a professional data entry company under contract. Alternatively, individual data entry operators might be sourced and brought in to work at the assessment centre under the direction of the data manager. In choosing which model to use, the advantages and constraints around employment law and contract procurement regulations should be considered.
3.7 Physical infrastructure

A high-quality assessment program will be supported by a high-quality physical infrastructure that is well-fitted to the tasks to be undertaken. The quality of the infrastructure will depend on the availability of resources and the budget. For this reason, infrastructure needs should be foreseen and planned at the beginning of the project, at or before the time of budget planning. This section gives a very broad overview of the types of infrastructure resources that should be considered. When budgeting, experts in the relevant fields should be consulted about specifics – for example, information technology experts about network requirements or statisticians about statistical software.

3.7.1 IT infrastructure

Physical (or virtual) information technology infrastructure may include:

- servers
- networked hard drives for shared access and storage
- internet connection
- cloud access/storage
- personal computers
- a project website.

3.7.2 Software

Generic and specialist software is required. Software may include:

- security software, including antivirus and firewall software
- database software for item banking
- editing software
- desktop publishing package
- statistical analysis package
- Item Response Theory (IRT) scaling software.

There may be open source licences available for some software requirements. A specialist should be consulted to identify appropriate software options. If open source software is considered, the specialist will need to be sure that the appropriate support is available and affordable.

3.7.3 Miscellaneous physical infrastructure

Other physical infrastructure that needs to be budgeted for includes:

- a printer
- a scanner or photocopier
- telephones, including a dedicated project hotline
- a storage site for assessment materials – should be secure from theft, fire and flood
- a secure space for conducting the field trial and main survey coding operations.
3.8 Summary

Ensuring that there are sufficient human and physical resources is important at the outset of an assessment program. The structure of an assessment team will be different for every assessment program. However, for every program it is important to ensure that the team has the right mix of skills and expertise and that the responsibilities of the different team members are clearly defined. Consideration should also be given to the capacity development requirements of team members. If the expertise required is unavailable internally, outsourcing/insourcing certain roles may need to be considered.

This chapter has also discussed the need for assessment programs to be supported by high-quality physical infrastructure. While the particular team and infrastructure needs will vary depending on the assessment program, it is always essential that resource requirements are established and budgeted for as early as possible to ensure programs are of high quality.
## Checklist: Project team and infrastructure

If you are planning or implementing a large-scale assessment, have these points been addressed?

### Human resources

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Has the organisation(s) that will design and implement the assessment program been identified?</td>
</tr>
<tr>
<td>2.</td>
<td>Has a project manager or project coordinator been appointed?</td>
</tr>
<tr>
<td>3.</td>
<td>Has a project team structure been planned and documented?</td>
</tr>
<tr>
<td>4.</td>
<td>Have senior management roles been appointed?</td>
</tr>
</tbody>
</table>

### Physical resources

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5.</td>
<td>Has the necessary IT infrastructure been identified?</td>
</tr>
<tr>
<td>6.</td>
<td>Is the necessary IT infrastructure in place or is the necessary IT infrastructure planned?</td>
</tr>
<tr>
<td>7.</td>
<td>Has the necessary software been identified and obtained?</td>
</tr>
<tr>
<td>8.</td>
<td>Have the miscellaneous resources been identified and obtained?</td>
</tr>
</tbody>
</table>

### Design and implementation issues

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9.</td>
<td>Is there adequate budget provision for human resources?</td>
</tr>
<tr>
<td>10.</td>
<td>Is there adequate budget provision for physical resources?</td>
</tr>
<tr>
<td>11.</td>
<td>Have outsourcing and insourcing options been identified?</td>
</tr>
<tr>
<td>12.</td>
<td>Have companies and individuals been identified for outsourcing and insourcing options?</td>
</tr>
<tr>
<td>13.</td>
<td>Is there adequate budget provision for outsourcing and insourcing options?</td>
</tr>
<tr>
<td><strong>Glossary</strong></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td><strong>Assessment framework</strong></td>
<td>A formal document that describes and discusses what an assessment intends to measure. The assessment framework explains the purpose and design of the assessment. See Chapter 5.</td>
</tr>
<tr>
<td><strong>Cloud access</strong></td>
<td>The ‘cloud’ is a term to describe a networked set of data centres. Among other things, access to the cloud allows the user to store and share large-scale datasets.</td>
</tr>
<tr>
<td><strong>Codebook</strong></td>
<td>A documentation of characteristics of the item that are needed at the time of analysis. This information includes a unique identifier, the domain or subject that the item is measuring, and the correct answer.</td>
</tr>
<tr>
<td><strong>Contextual framework</strong></td>
<td>A formal documentation, often within the assessment framework, of why and how characteristics of the test-taker or the test-taker’s environment are to be measured. See Chapter 5.</td>
</tr>
<tr>
<td><strong>Gap analysis</strong></td>
<td>Identifying the difference between current human and physical capacity and the capacity needed to achieve high-standard business processes – such as, in the context of this handbook, to achieve a high-quality educational assessment program.</td>
</tr>
<tr>
<td><strong>Item Response Theory (IRT) scaling software</strong></td>
<td>Software that is used to transform a test-taker’s raw test scores onto a scale that accurately and reliably expresses the test-taker’s ability. See Chapter 12.</td>
</tr>
<tr>
<td><strong>Items</strong></td>
<td>The questions or tasks used in an assessment.</td>
</tr>
<tr>
<td><strong>Psychometrics</strong></td>
<td>Theory and methods of measuring psychological traits, such as mathematical ability or motivation to read.</td>
</tr>
<tr>
<td><strong>Test developers</strong></td>
<td>Those responsible for producing test content, including contributing to the test design, writing items, helping to interpret trial data to select items for the main survey, and interpreting main survey data for reporting.</td>
</tr>
<tr>
<td><strong>Train-the-trainer model</strong></td>
<td>A system where a group of people are taught to provide training to others. Used when there are large numbers of people to train and/or they are geographically dispersed.</td>
</tr>
<tr>
<td><strong>Webinar</strong></td>
<td>A seminar facilitated via the internet. Often called web conferencing and similar to video conferencing.</td>
</tr>
</tbody>
</table>
Technical standards are guidelines that establish the qualities or requirements of a given process or output.

In the context of large-scale assessments, technical standards are designed to assist the implementing agency and relevant stakeholders of an assessment program. They assist by explicitly indicating the expectations of data quality and program implementation endorsed by the governing body of the program (e.g., the program steering committee). They also set the timelines of the activities involved. The standards formulate levels of attainment, while more detailed implementation procedures are defined in the operations manuals (as discussed in Chapter 10).

The construction of technical standards is one recognized way of contributing to the quality of an assessment. Documenting the adherence to, and deviation from, the technical standards will improve the confidence of policymakers and planners in the results and help make the program resistant to technical criticism.

This chapter presents a set of technical standards that are applicable to large-scale assessments and may be appropriate in the Indian context. This chapter also provides examples of specific technical standards that are adopted in various international, regional, national, and state assessment programs. These examples of best practice can be used as guidelines for future assessment programs, or to strengthen existing programs in India.

It is important to note that the standards discussed in this chapter relate primarily to large-scale assessments. Although it may be conceptually valid to apply technical standards to a wide range of assessments and testing processes, the scope of this chapter aligns with the overall scope of this handbook.

A broader definition of a ‘standard’ is a statement or set of statements that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes, and services are fit for their purpose (see http://www.iso.org/iso/home/standards.htm).
4.1 Purpose and types of technical standards

Technical standards set a benchmark of best practice for the evaluation of the assessment instruments and the assessment process as a whole. According to the Standards for Educational and Psychological Testing [1], there are three categories of standards:

- **Primary standards** need to be met before the instruments can be used.
- **Secondary standards** need to be met unless they are not feasible in situations beyond the control of the developer and user.
- **Conditional standards** have varying levels of importance depending on the situation and setting where the standards apply.

The designation of any particular standard, however, is context-specific and purpose-dependent [1]. For example, primary standards in one assessment program in a European country may be designated as secondary in a different assessment program in India.

Given the scope of this chapter, it is also important to reiterate that the standards need to be established around the set of program components or tasks that relate to the major processes, milestones and deliverables specific to an assessment program. In major assessment programs [2], the components can be broadly categorised into at least two groups that relate to (i) the data aspect and (ii) the management aspect of the program.

Given below is a list of the major components that are common across state, national, regional and international assessments. Their composition, however, may vary depending on the program. Each component will be discussed in more detail in the following sections within this chapter:

- field trialling standards
- sampling standards
- data standards
- psychometric standards.
4.2 Field trialling standards

Field trials have two important purposes: (i) to give the assessment team an opportunity to try out the logistics of their test procedures, and (ii) to allow for detailed psychometric analyses of the items so that only those items that are suitable are included in the main survey.

A field trial is administered in students' language of instruction. According to international best practice, a field trial should occur in an assessment language if that language group represents more than 5% of the target population. In large-scale assessment, this standard has substantial implications for translation costs and logistics. As such, cost considerations need to be taken into account in setting context-specific standards regarding the number of language groups to be catered in the assessment project.

The international best practice regarding the psychometric purpose of the field trial establishes the standard that the field trial sample should include at least 200 students per test item, in each subject and each language.4

4These are international standards for languages that apply to more than 50% of the target population. For assessment languages that apply to between 5 and 50% of the target population, a different standard may be set (for example, the Programme for International Student Assessment (PISA) has set the field trial sampling standards for psychometric purposes at a minimum of 100 students per item).
4.3 Sampling standards

The following sampling standards are based on large-scale, multilingual, international studies [2] and are provided here as an example of international best practice. These standards apply only to the main survey and not to a field trial. Adopted sampling standards need to be set in place in the appropriate sampling or operation manuals specific to each assessment program.

4.3.1 Sample sizes

The sample size is dependent on the study design and can vary widely such that no set values can be given as guide (see Chapter 9 for more on this topic). However, it is best practice to set the total of combined school-level exclusions and within-school exclusions within each state and territory to no greater than 5% of the national target population5 within the respective state or territory.

After the study design has been set and the exclusion criteria have been identified, the sampling standards need to be established with the aim of achieving a level of precision comparable with international studies such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA).

It is international best practice that sample sizes should be large enough to yield a 95% confidence interval for all relevant estimates. In a simple random sample, this degree of statistical precision translates to a minimum of 400 students for student-level estimates. More complex sampling and analysis designs require compensation for various design effects such as clustering and subgroup analysis to achieve the same sampling precision (see [3] for more technical details).

4.3.2 School response rates

Best practice is that the school response rate is at least 85% of sampled schools. If a response rate is below 85% then an acceptable response rate can still be achieved through a systematic use of replacement schools.

5The national target population is defined as the population to which inferences will be made.
4.3.3 Student response rates

Best practice is that the student response rate is at least 80% of all sampled students across responding schools.

4.3.4 Sample coverage

Best practice is that the defined target population covers 95% or more of the desired target population (see Chapter 9 for definitions of sampling terminology). That is, school-level exclusions and within-school exclusions combined do not exceed 5%. Exclusions are predefined criteria for schools and students that will be excluded from the assessment (see Section 9.2 for more details), for reasons that may be due to geographical inaccessibility, logistical considerations, very small schools, and within-school exclusions [2].

Technical standards should specify within-school inclusion criteria that are applicable in the Indian context, including certain types of disability and language limitations. Disability may include functional, emotional, or intellectual disability which severely restricts the test-taker’s ability to perform in the assessment situation. The within-school language exclusion relates to students who are unable to read, speak, or otherwise overcome the language medium barrier in the assessment situation.

6Current practice in the NAS is a combined exclusion rate in the range of 15–20% across the various NAS classes.
4.4 Data standards

4.4.1 Data entry standards

Data entry standards relate to manual data processing as well as the coding and scoring process. Coding is the process of placing numerical labels on student responses so as to capture qualitative distinctions between those responses. Later in the analysis, the codes are scored to reflect an intended ordering (in terms of merit) that was assigned to the different performances that were coded for qualitative distinctions. Coding schemes and scoring procedures are specified in the operational manuals. The quality assurance component of data entry is part of the technical standards that need to be set. This includes indices of inter-rater agreement, recruitment and training standards, and data entry audit standards.

There are two common types of data entry errors: transcription error and transposition error. Transcription errors occur when a character in the data is incorrectly entered as another character in the dataset. This error is common in manual data entry but can also occur in digitally scanned data, especially when the scanned data is ambiguous (e.g., numbers that might be confused with letters such as zero=0 written as an uppercase letter=O, scanner defects, or poorly written characters). In the vast majority of cases, transposition error is a human error, and occurs when the data entry operator switches the order of a series of characters (e.g., 123 is entered as 132).

International best practices for minimising data entry errors include:

- utilising automated data validation procedures
- double entry (including double scanning for OMR/OCR processes)
- coding verification – the data entry procedures need to be reviewed if the following error rates are exceeded: maximum 5% error rate for automated systems, 10% error rate for human operators [4].

4.4.2 Test administration standards

Test administration standards set the procedures that need to be followed during the test sessions. These procedures are specified in the field operations manuals and relate to:

- student sampling
- test session timing
- maintaining test conditions
- rotation of booklets (e.g., ensuring booklets are evenly distributed)
- seating arrangements
- time requirements (e.g., whether the test has time constraints or not)
- student tracking (e.g., procedures for filling out tracking forms or field notes)
- batching (e.g., compiling assessment instruments into groups or packages for processing)
- random assignment of booklets
- quality control (including manual check of test materials).

---

7These refer to statistics (e.g., Cohen’s kappa, Fleiss’ kappa, inter-rater correlation, Krippendorf’s alpha, etc.) on the degree of agreement or consensus between two or more raters.

8The acceptable level of error rates is dependent on the program. High-stakes programs may require 100% data entry accuracy.
It is good practice that field investigators are trained in person, with the training conducted as close to the level at which the test administration is conducted. For example, if the test is conducted at a classroom level, it is best to train the field investigators collectively as close to the classroom level as possible (e.g., at a school level). Train-the-trainer models can work, but the further removed the training is from the actual site and the more levels the training process has to go through, the less effective the training becomes to the end user and the greater the information loss in the transmission of knowledge.

It is important to ensure that the relationship between field investigators and participating students does not compromise the credibility of the test session. It is essential that the independence of the field investigators from the students and school staff is maintained. High-stakes testing may require this independence of field investigators from other parties that may be indirectly affected by the results of the assessments (such as state or regional directors of school systems, teacher union staff, curriculum authorities, and so on).

Test administration and field operations require regular and independent monitoring for quality assurance purposes. International best practices require at least 15% of test administration sessions at the state level to be independently monitored. As the logistics of independent monitoring entails financial costs, the percentage of test administration sessions that are independently monitored can be reduced to 10% provided that the monitoring remains independent and the selection of administrations to be monitored remain random. It is international best practice to document the extent and standard of compliance of the observed test administrations with the prescribed procedures with appropriate monitoring formats [5].

The standards on minimum educational and professional qualifications of field investigators depend on the program and have direct implications on costs of hiring field investigators, as more stringent standards would require staff to have higher qualifications. Cost considerations also need to be taken into account when planning for test administration training workshops.

### 4.4.3 Test security standards

Standards must be set by the assessment team and followed by all the other participants in an assessment program that are directly involved (including subcontractors) in the implementation of the program and analysis of the data in order to ensure that program materials designated as secure are kept confidential at all times. Secure materials include all test materials, data and draft materials. The scope of the security standards includes all levels of the assessment process, from the classroom all the way to state or national levels.

The security standards should cover access protocols and security procedures. Access protocols include issues such as: Who are the approved staff and participants? What aspects of the program are accessible to specific types of staff and participants? Security procedures, such as confidentiality agreements, security audits and monitoring, need to be set in place before the secure assessment materials are released. It is also good practice to set contingency standards in case security breaches occur.
4.5 Psychometric standards

Psychometric standards quantitatively define the technical quality of the test data and the interpretations that can be derived from the test results. There is extensive literature on psychometric standards [1, 2, 6, 7] but specific psychometric standards and guidelines on their implementation are often program-dependent. Psychometrics is a complex and comprehensive topic: this section provides only an overview. The key issues discussed here relate to reliability, validity, important indicators of item characteristics under Item Response Theory (IRT), and differential item functioning (DIF).

4.5.1 Reliability

In Classical Test Theory (CTT), there are several types of reliability estimates [see 6 and 7 for more details on different types of reliability estimates]. The most common type used in large-scale testing is internal consistency. Internal consistency as a type of reliability estimate assumes that the test is unidimensional, or measuring a single construct. Internal consistency is measured by Cronbach’s alpha. Reliability standards typically require values of Cronbach’s alpha to be no less than 0.8 for high-stakes testing individual assessments and 0.6 for group assessments [8]. However, while high alpha values indicate higher internal consistency among the test items, extremely high values (e.g., approaching 1.0) may be an indicator that the items are redundant.

In Item Response Theory (IRT), the reliability coefficient is linked with the separation index.\(^9\) The separation index is an indicator of how sensitive an instrument is in differentiating (or ‘separating’) between any given number of levels (e.g., ability levels). If a test is required to separate only two levels (i.e., low and high performers) the corresponding separation reliability is 0.80; a separation reliability of 0.96 indicates that a test is able to separate up to five levels [10].

4.5.2 Validity

Validity needs to be established to support the use of any assessment instrument and interpretation of the results of the assessment process [8]. From a technical standards perspective, establishing validity requires gathering evidence which can be grouped into the following major sources [1]:

- Evidence based on test content, including how well the test covers the subject and how relevant the content is to the subject.
- Evidence based on response processes of test-takers, which can involve checking that the students’ cognitive processes during the test match the construct being measured [1], for example through cognitive laboratories. This type of evidence also includes evidence from judges or markers who may be interpreting the responses (for example in open-ended items).
- Evidence based on the internal structure of the test, in a way that the test structure matches the theoretical structure of the construct being measured. This type of evidence relates to the test framework, dimensionality (i.e., the test measures only one dimension or construct), and relationships among the test components.
- Evidence based on the relationship between the test and other variables that are external (but not construct-irrelevant) to the test, for example between students’ National Achievement Survey Class X scores in English and their Examination Board scores in English.

\(^9\) Separation index is defined as the ‘ratio of the unbiased estimate of the sample standard deviation to the root mean square measurement error of the sample’ [10, p. 162]. This ratio indicates how large the model error variance is in proportion to the true variance. For more details on the separation index, see [9].
4.5.3 Item difficulty

In Item Response Theory (IRT), an item provides maximum information on the students who have latent abilities that match the item difficulty (see Chapter 12). This concept of targeting the test to the test-takers means that the range of item difficulties of a test should match the range of ability of the target test-takers. It is in line with international best practice that there should be enough difficult items at the top end as well as enough easy items at the bottom end of the scale to target the students across the entire range of ability distribution and provide reliable information on the measured constructs.

4.5.4 Item discrimination

In Classical Test Theory (CTT)\textsuperscript{10}, item-rest and item-total correlations are related to the concept of item discrimination. Item-rest and item-total correlation coefficients provide an index of the strength of relationship between an item and the measurement scale that it belongs to. As such, the correlation coefficients measure the discriminating power of an item with respect to the measurement scale (which may or may not include the item for item-total and item-rest correlations respectively). International best practice is to retain only those items with item-rest and item-total correlations that are not less than 0.2. This ensures that items which do not have at least a moderately strong relationship to the scale are not included in the item pool.

4.5.5 Fit statistics

In constructing measurement models, it is important to check how well the response data from the test-takers fit the measurement model that is being used to analyse the results. Fit statistics are indicators of model fit\textsuperscript{11} for both person data (ie a student’s response pattern) and item data (pattern of responses to an item).

Under the one-parameter logistic model, the item and person parameters are interchangeable and therefore the fit criteria for both are the same [11]. (The one-parameter model is also referred to as the IPL or Rasch model. See Section 12.1.2 for details). However, because item parameters are expected to be relatively more stable than person parameters, standards on fit statistics are often set on item fit rather than person fit. It is international best practice to use a weighted mean-square fit statistic and a reasonable range of 0.8 to 1.2 as indicator that an item does not deviate significantly from the model.\textsuperscript{12}

4.5.6 Differential item functioning

For an item to measure an ability (eg mathematics ability) in a fair manner, people of the same ability should have the same probability of getting the answer right. When people with the same ability from different groups (eg boys and girls) have different probability of successfully answering an item, that item is said to be functioning differentially across groups. It is important that assessments are designed to minimise differential item functioning (DIF).

\textsuperscript{10}Item discrimination under an IRT perspective, specifically in two-parameter logistic and three-parameter logistic models, is a function of the item characteristic curves. The derivative of this function, representing the slope of the item characteristic curve, reaches a maximum value at the inflection point of the curve. This maximum is reported as the discrimination parameter estimate in the two-parameter logistic and three-parameter logistic models.

\textsuperscript{11}That is, the extent that a response pattern distorts the measurement model.

\textsuperscript{12}NAS Class III, V and VIII use a two-parameter logistic model. The recommendations based on international best practice regarding the range of reasonable fit statistics still apply.
There are a number of methods to statistically analyse DIF and therefore guidelines as to what might be considered non-negligible DIF vary depending on the method. It is, however, good practice that DIF analysis be conducted whenever there are reasonable concerns that subgroups within the target population are affected by factors that are not related to ability (ie construct-irrelevant variance) [1]. It is also good practice to include relevant results from a DIF analysis in reporting the overall findings of any assessment project.
4.6 Summary

This chapter outlines key issues that need to be considered in setting technical standards for large-scale assessment programs. The standards provided here are based on international best practices and cover both the data-related and management-related aspects of an assessment program.

The scope of a set of technical standards and its level of detail is dependent on the program. Large-scale international assessment programs may have a very comprehensive set of technical standards while small regional assessments may have a set of technical standards that just covers a limited number of assessment tasks.

Regardless of program scale, it is recommended that technical standards be set before the assessment process is implemented. It is also recommended that once standards have been set, protocols for non-compliance are set and instances of variations from the standards documented.
# Checklist: Technical standards

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Purpose of technical standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have technical standards been set according to context-specific and purpose-dependent considerations of the assessment program? [ ]</td>
</tr>
<tr>
<td>2. Have technical standards been set for both the data-related and management-related aspects of the program? [ ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field trialling standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Have technical standards been set on the sample size and coverage of language groups for the field trial? [ ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sampling standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Have standards been set on sample size, response rates and sample coverage? [ ]</td>
</tr>
<tr>
<td>5. Are the exclusion criteria for school and within-school levels clear and well-defined? [ ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sampling standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. Have the data entry standards been set? [ ]</td>
</tr>
<tr>
<td>7. Have the test administration and test security standards been set? [ ]</td>
</tr>
<tr>
<td>8. Are the data standards appropriate and achievable in the contexts of the program? [ ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychometric standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Have the psychometric standards been set? [ ]</td>
</tr>
<tr>
<td>10. Are the psychometric standards based on international best practices while also taking into consideration the context of the program? [ ]</td>
</tr>
</tbody>
</table>
### Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive laboratories</td>
<td>In assessment, cognitive laboratories are also known as cognitive interviews. These involve settings where students are observed and studied in detail to investigate the thinking processes that they employ when performing assessment tasks.</td>
</tr>
<tr>
<td>Confidence interval</td>
<td>An interval that specifies a range of values for a parameter estimate, based on a predefined confidence level, and calculated from one sample of the population. The confidence level (usually 95%) for an interval indicates the proportion of intervals, computed from all possible samples, that includes the true value of the parameter being estimated.</td>
</tr>
<tr>
<td>Design effect</td>
<td>An adjustment used in complex survey methods to assist in determining the study’s sample size.</td>
</tr>
<tr>
<td>Item-rest correlation</td>
<td>Correlation between each item and the scale, not including the item being correlated.</td>
</tr>
<tr>
<td>Item-total correlation</td>
<td>Correlation between each item and the scale in its entirety (including the item itself).</td>
</tr>
<tr>
<td>Simple random sample</td>
<td>A method of sampling where every member of a population has an equal chance of being selected.</td>
</tr>
<tr>
<td>Excluded population</td>
<td>The population that does not form part of the defined population.</td>
</tr>
</tbody>
</table>
References


Assessment framework
Assessment framework

The development of an assessment framework is often an iterative process – one that takes place step-by-step.

This chapter describes assessment frameworks and their purposes. The first part of the chapter is about frameworks for assessing areas of learning. The second part is about frameworks for collecting contextual information about the students taking an assessment. The last part of the chapter is about how an assessment framework document can be organised.

Ideally, work on a framework begins before any assessment items are written, so that the framework is available in draft form as a guide to test development. However, it is also important that a framework is seen as a work in progress: flexible enough to be modified in the light of practice and the outcomes of at least the early administrations of the assessment. Those contributing to the development of a framework may include academic experts, curriculum authorities (for a curriculum-based test) and assessment and evaluation experts. It is valuable to bring together people with different perspectives, to draw on a wide range of informed opinion. These strategies will help to ensure that the framework has authority and wide acceptance.
What is an assessment framework?

An assessment framework is a description and discussion about what an assessment intends to measure. A framework helps the people who are interested in the assessment to understand what the assessment is about. It also helps them to understand what the assessment results mean.

The framework includes a definition of the domain – the area of learning that is being measured. It explains which aspects of the domain are included in the assessment. The framework also outlines how the assessment is constructed, by setting out the proportion of items for each aspect of the learning domain that make up the assessment. It describes the response formats that are used, and the length and number of items in the assessment. The framework may also outline how the results of the assessment will be reported.
5.2 What are the purposes of an assessment framework?

There are several reasons for developing an assessment framework. One important reason is that it underpins the validity of the assessment, by making explicit what the aim of the assessment is, and what it will cover in terms of content, skills and knowledge.

Another purpose of the framework is to define terms relating to the assessment. This means that when people discuss the assessment they can communicate clearly and without misunderstanding.

If the assessment is to monitor trends, the framework helps to ensure that the test is measuring the domain and its context in a consistent way, over time. For example, the proportion of items measuring each aspect of the content is documented in the framework. This is a guide to ensure that the balance of aspects remains the same from one year to the next. Sometimes it is necessary to change an assessment: for example, a new aspect of a domain might be added to the curriculum, and the new aspect needs to be reflected in the assessment. The framework needs to describe the change, and explain why the change is being made. In this way, there is a record to help interpret any change in results.

For test developers, the assessment framework is especially important because it provides a guide for item writing, by explaining what areas the test must cover, how the areas are defined, and what proportion of the test should focus on each area. At a later point the framework can be used to review how well the assessment instrument has met its purpose.

Ideally an assessment framework is developed collaboratively, involving input and feedback from stakeholders such as education agencies and teachers, domain experts and item writers. Those involved in developing an assessment framework feel some ownership of the assessment. This increases the likelihood that the results of the assessment will be taken seriously, and ultimately will have an impact on improving learning.
5.3 Framework for a learning domain

Some assessments, especially assessments for school children, are based on a curriculum, and often on just one class (grade) level of the curriculum. In such cases, much of the content of the assessment is predetermined, and the definitions and descriptions of the learning domains and its parts are already clearly articulated. Other assessments are competency based: that is, they are designed to measure more generic knowledge and skills that are cross-curricular or that go beyond any particular school curriculum.

Competency-based assessments generally need more detailed frameworks, because they do not refer to any pre-existing design or set of definitions or descriptions. However, even when an assessment is based on a clearly articulated curriculum, what is assessable in a large-scale assessment is unlikely to cover all elements of the curriculum. Accordingly, assessment frameworks are needed for both curriculum- and competency-based assessments, to specify what is and what is not covered by the assessment.

The essential components of any learning assessment framework are:

- an explanation of the purpose of the assessment, and who is being assessed
- a definition and description of the domain
- an outline of the content areas of the domain that are included in the assessment
- an outline of the skills that are intended to be measured in the assessment
- specifications for the design and for operationalising the assessment, including:
  - the proportion of items focusing on each area of content and skill
  - information about test delivery medium and response formats
  - a statement of the length of the test and the number of items.

A brief description of each of these components follows.

5.3.1 Assessment purposes and who is being assessed

A preamble or overview of the assessment program should be provided. This may be an elaborate introduction, including such elements as the history of the development of the program, and a description of who has contributed to its formulation [eg 1, 2]. However, at a minimum it should state the aim of the assessment: for example, ‘accreditation for individuals at the end of a stage of schooling’, ‘a report on a system’s performance’, or ‘monitoring progress of a cohort over time’.

The target population should also be defined, according to jurisdiction (for example, national, state or a smaller entity) and whether by class or age. An example of a succinct statement of purpose is provided in the National Achievement Survey (NAS) Class X Assessment Framework [3], as shown in Exhibit 4.

Exhibit 4
Statement of NAS Class X assessment purpose

<table>
<thead>
<tr>
<th>Purpose of the Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>The achievement survey at secondary stage (Class X) aims:</td>
</tr>
<tr>
<td>- To study the achievement levels of students of Class X in a Modern Indian Languages, English, Mathematics, Science and Social Science.</td>
</tr>
<tr>
<td>- To study the difference in achievement levels with regards to area, gender, social group, board and management of schools.</td>
</tr>
<tr>
<td>- To study the effect of intervening variables like home, school and teacher on students’ achievement levels.</td>
</tr>
</tbody>
</table>
5.3.2 Definition and description of the domain

It is important for the framework to be explicit about what is being assessed. Just naming the domain, or subject, is not sufficient. Conceptual understandings of even apparently well-known domains differ, and change. For example, as Anderson and Morgan point out, definitions of reading proficiency have varied over time and across education systems, from simply being able to pronounce words, to the ability to explain the meaning of words, to comprehending texts and using information from texts, extending even to enjoying and appreciating texts [4].

Some frameworks provide one or two sentences to summarise the domain that lies at the heart of the assessment. The Programme for International Student Assessment (PISA) 2012 definition of reading literacy, for example, is, ‘an individual’s capacity to understand, use, reflect on and engage with written texts, in order to achieve one’s goals, to develop one’s knowledge and potential, and to participate in society’ [2]. In other cases, where an assessment is firmly curriculum based, reference to the curriculum document’s definition of the domain may be sufficient.

5.3.3 Outline of the domain content areas

This part of an assessment framework lists and describes the content categories (or sub-domains) that are to be included in the test. These content areas are specific to the domain, for example, typical content areas in science are biology, physics and chemistry. In reading, content areas are often represented as text types: narration, information, persuasion, and so on. In mathematics, typical content categories are number, space, measurement and statistics.

The content categories may vary according to age or class level. For example, in the Trends in International Mathematics and Science Study (TIMSS), the Grade 4 mathematics content domains are number, geometric shapes and measures, and data display, whereas for Grade 9 they are number, algebra, geometry, and data and chance [1]. More or less explanation of these categories may be needed, depending on whether they are new areas or well-defined in a curriculum document.

Both the scope and the limitations of what is to be assessed need to be clearly outlined. For example, the United Kingdom’s Key Stage 2 assessment frameworks for English grammar, punctuation and spelling, which are based on the English National Curriculum, include sections on elements of the curriculum that cannot be assessed fully in the Key Stage 2 national assessment [5, 6].

5.3.4 Outline of the skills to be measured

Skills – sometimes called ‘processes’, ‘cognitive domains’ or ‘aspects’ – are the ways of thinking or intellectual approaches that develop as individuals become increasingly proficient in a domain. Similar to the content areas of the domain, the skill areas to be addressed in the assessment need to be outlined and explained in the framework. Skills may be defined in a generic way, using a broad set of cognitive approaches, like that in the NAS Class X assessment framework’s description of skills for science and social sciences [3] shown in Exhibit 5.
In other assessments, skills are defined in a way that is quite domain specific. For example, writing is one of the domains in the South East Asian Primary Learning Metrics program. Five skill areas have been identified as fundamental to writing literacy: generating ideas; controlling text structure and organisation; managing coherence; using vocabulary; and controlling syntax and grammar [7].

5.3.5 Specifications for designing and operationalising the assessment

The preceding part of this chapter has indicated that an assessment framework should provide a definition of the domain, and should describe the content and skills to be included in the assessment. The framework must also explain how these concepts will be operationalised in the items.

5.3.5.1 Blueprint for content and skills

Part of the process of operationalising the assessment involves determining the proportion of items in each content and skill area. This is sometimes called the assessment blueprint. The proportions assigned represent the relative importance of each variable category. For example, in the skill categories of knowing, applying and reasoning in the NAS Class X assessment (Exhibit 5), the recommended distribution of items is in the ratio 1:2:1. This suggests that applying is judged to be the most important of the skills in the subjects of science and social science at Class X.
The proportions assigned to different content or skill areas may differ across class and age levels, even when the categories themselves are the same. An example of this is found in the American National Assessment of Educational Progress (NAEP) reading assessment framework for 2013 [8], as shown in Exhibit 6.

In the NAEP reading framework, the content (text type) areas have different proportions for Grades 4, 8 and 12. For example, there is a larger proportion of items about informational texts in the Grade 8 assessment than in the Grade 4 assessment. There is an even larger proportion in the Grade 12 assessment. This implies that reading informational texts becomes more important as students advance through the grades.

**Exhibit 6:** Percentage distribution of text types in NAEP at three grade levels

<table>
<thead>
<tr>
<th>Grade</th>
<th>Literary</th>
<th>Informational</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>70</td>
</tr>
</tbody>
</table>

In some assessments, the proportions of items for each area are given as ranges, not single numbers. This is especially useful in the early stages of an assessment’s development, to allow for adjustments to be made based on the experience of implementing the assessment. For example, an Australian test of literacy and numeracy, which will be administered to all initial teacher education students prior to graduation from 2016, uses target ranges in its assessment framework [9]. This assessment framework provides the example of target ranges shown in Exhibit 7. The ranges allow some flexibility in constructing the test in its early stages of implementation.

**Exhibit 7:** Target ranges for processes (skills) from a numeracy assessment framework for initial teacher education students

<table>
<thead>
<tr>
<th>Process</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifying mathematical information and meaning in activities and texts</td>
<td>15–25%</td>
</tr>
<tr>
<td>Using and applying mathematical knowledge and problem-solving processes</td>
<td>50–60%</td>
</tr>
<tr>
<td>Interpreting, evaluating, communicating and representing mathematics</td>
<td>20–30%</td>
</tr>
</tbody>
</table>
5.3.5.2 Test delivery medium and response formats

An assessment framework should specify whether the assessment will be delivered via paper-and-pen or via computer. It should also specify the formats in which students will need to respond. (Detailed information on different response formats is provided in Chapter 6.) Will the assessment require students to respond only to multiple-choice items? Will they need to construct short answers? Will they need to write essays? Or will the assessment require students to respond in a variety of formats?

To decide on response formats, test developers need to think about the nature of the domain and which formats can validly capture information about the students’ knowledge, skills and understanding. To take an extreme case, the results of an assessment of dance performance that was only made up of multiple-choice items would not be taken seriously. As the assessment framework for the Progress in International Reading Literacy Study (PIRLS) says, the decision about response format ‘is based on the process being assessed, and on which format best enables test-takers to demonstrate their reading comprehension’ [10, p. 67]. That said, in large-scale paper-and-pen assessments the range of response formats is limited by practicality. Multiple-choice items and short constructed response items (where students write a short answer) are the most manageable formats, in terms of both cost and reliability of marking.

The framework should not just specify the types of response format used in the assessment, but also the proportions of each type, where there is more than one. Again, this may be expressed in terms of target ranges, as shown in Exhibit 7.

5.3.5.3 The length of the test and the number of items

The assessment framework should include a statement about the length of time students will be given to answer the set of items, and how many items will be in each test. These pieces of information are important for test-takers and administrators of the tests, as well as for item writers.

5.3.6 Additional components of an assessment framework

Some assessment frameworks include a section on how the results of the assessment will be reported. Even if reporting is not explicitly discussed in a framework, it must be considered when the framework is being developed, because what is to be reported will influence how the test is shaped. For example, if a requirement in a test of mathematics is to report on different sub-domains (such as algebra and geometry), sufficient items must be included on each sub-domain to allow meaningful results to be obtained for reporting purposes.

Another area that is often included in an assessment framework is an outline of the contextual information to be collected to support the interpretation of the cognitive results. This area is discussed in more detail in the following section of this chapter.
5.4 Contextual framework

Assessment frameworks generally incorporate a contextual component, referred to as a contextual framework. The contextual framework acts as the foundation for contextual information to be collected as part of the survey. It establishes what contextual information is to be collected, as well as the justification for doing so.

Given that much of the variation that exists in student outcome data can be explained by the various contexts that surround the student, the contextual framework is an important component in large-scale research. It can be used to help discover the meaning behind differences in observed test outcomes, and contribute to the development of strategies to improve such outcomes in students. The contextual framework should define what information needs to be collected in the contextual questionnaires, which can be designed for respondents at the student, parent, teacher, school, regional and national levels. (See Chapter 8 for information on developing contextual questionnaires.)

5.4.1 Contextual factors

The development of the contextual framework should follow a well-considered approach. As a starting point, prior research should be used as a guide as to what contextual influences would be expected to influence student outcomes. For example, a numeracy assessment may wish to include a measure of student engagement with mathematics given the suggested link between engagement and achievement in this area [11]. It is worthwhile consulting the framework documents and/or contextual questionnaires for other studies with similar aims. Studies conducted in the Indian context would be most suitable.

There may be a desire to also include factors that may not necessarily be expected to influence student outcomes, but for which the data are of interest. Whatever areas are chosen to be included in the contextual framework, they should be well-defined and included with a clear purpose in mind that reflects the aims and goals of the program. (See Chapter 2 for a discussion on assessment program goals.) Deciding on what information should be defined in the contextual framework (and thus included in contextual questionnaires) should be done with extensive consultation with all major stakeholders in the project.

A common structure for presenting the range of contextual factors included within a contextual framework is to use a multi-level design that distinguishes between the broad types of contextual factors that are typically included in contextual questionnaires [13, 14, 15]. These levels include factors relating to:

- the individual context of the student
- the home environment of the student
- the school and classroom level
- the wider community.

Where possible, justification for the inclusion of the contextual factors in the framework should be included, citing past research relevant to the area.
5.4.1.1 Individual context

This level relates to factors at the individual student level that might be expected to be related to outcome variables. The most common example of these is basic student characteristics such as the age and gender of the student. Other factors of interest at the individual student level may relate to students’ attitudes, perceptions, behaviours and self-concept. For example, the NAS asked questions relating to activities that students participate in outside of school, students’ likes and dislikes of school, and the subjects they study at school [12].

5.4.1.2 Home environment context

The role that home and family contexts play in the development of students’ knowledge attitudes, beliefs and behaviours is substantial. These include information relating to family members, language spoken at home, ethnic background, social class, home resources available for learning and the context of the home environment. Given the wealth of research that suggests a link between socio-economic background of students and achievement, measures of socio-economic status are almost always included in contextual questionnaires. Types of questions that measure aspects of socio-economic status include:

- physical composition of the house (e.g. what are the walls and roof made of, what type of lighting is used)
- household facilities (e.g. electricity, running water)
- household possessions (e.g. TV, computer)
- household income
- parents’ education
- parents’ occupation.

For example, in the NAS Class V program, a socio-economic index was produced based on data collected on questions relating to parents’ education, parents’ occupation, and household possessions [12].

5.4.1.3 School and classroom context

The environment of the school and classroom has a strong influence on the development of knowledge and competencies in young people. Information about school and classroom level factors can be collected using teacher and school questionnaires.

Typically school factors include characteristics of the school, organisational practices within the school, relationships between students, teachers, school heads and parents, as well as resources available to the school. Information may be collected about teachers at the school including their training and background. For example, in the NAS Class V assessment, the questionnaire asked teachers to indicate whether they participated in training programs based on a revised National Curriculum Framework [12]. Assessments often collect information about the respondents. Perspectives on each of the factors listed may be sought at the teacher level or school level (usually from the school head).

Information on teacher resources may be collected at the classroom level (for example, a question might ask about teachers’ access to a teaching learning material grant). Additional factors relating to teacher attitudes, perceptions and practices are commonly of interest. Information from the student questionnaire can also be used to gather perceptions of classroom and school practices.
5.4.1.4 Wider community context

Information for the wider community context relates to factors that might be expected to influence student outcomes from different levels. One level is the local community, for example the urbanisation of the local area or the interaction between the school and the local community. These factors would typically be included as part of a school questionnaire. The wider community context also may include factors at the region or national level, such as characteristics of the education system and population and geographical demographics. These would typically be sourced from external sources, such as from the Census of India website (http://censusindia.gov.in).
5.5 Organisation of an assessment framework

While well-regarded assessment programs have frameworks that include the features described in the previous section, they vary in their structure. Two examples of how all the different components are arranged in established large-scale assessments are shown in Exhibit 8 and Exhibit 9. Exhibit 8 shows extracts from the Table of Contents of the 2007 Writing framework of the NAEP program [8]. Chapters 1 to 3 of the NAEP writing assessment framework present the purpose of the assessment, the domain definition, and the outline of content and skills. Sections I and II cover specifications for the test and its design and implementation.

Exhibit 8:
Structure of the NAEP writing assessment framework, 2007

Chapter 1: Developing the Writing Framework
- The Nature of Writing and the NAEP Writing Assessment
- The Writing Process
- The Development Process for the Writing Framework

Chapter 2: Designing the NAEP Writing Assessment
- Rationale and Objectives for the Assessment

Chapter 3: Constructing and Scoring the Assessment
- Designing Topics

Section I: Assessment Specifications
- Content Specifications
- Technical Specifications
- Review Specifications
- Administration Specifications
- Scoring and Reporting Specifications

Section II: Task Specifications
- Format Specifications
- Scoring Rubric Specifications
- Reader Training Specifications
Exhibit 9 is an extract from the Table of Contents of the assessment framework for the international assessment program, PIRLS 2011 [10]. In the PIRLS framework, the purpose, definition, content and skills of reading are addressed in Chapters 1 and 2. Chapter 3 is dedicated to the context questionnaire framework. Test specifications are dealt with in Chapter 4.

**Exhibit 9:**
Structure of the PIRLS assessment framework, 2011

- **Chapter 1**
  - Overview of IEA's PIRLS Assessment
- **Chapter 2**
  - PIRLS Reading Purposes and Processes of Reading Comprehension
- **Chapter 3**
  - Contexts for Learning to Read
- **Chapter 4**
  - Assessment Design and Specifications
5.6 Summary

An assessment framework is essential for a robust and authoritative large-scale assessment program. It is a communication tool for the program, giving information to education agencies, school leaders, teachers, domain experts and the wider public. It provides information about the purpose, content and structure of an assessment. It supports the work of those involved in the design and implementation of the assessment. It gives technical guidance to test developers and supports a program’s stability over time.

Although stability in an ongoing program is important, an assessment framework should be regarded as an evolving document. It should be capable of adaptation to meet changing circumstances. Assessments often need to find a balance between continuity and change. An assessment framework can help to strike such a balance.
### Checklist: Assessment framework

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Process of developing an assessment framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Are test developers, data analysts and other evaluation and assessment experts involved in developing the framework?</td>
</tr>
<tr>
<td>2. Have stakeholders been consulted about the assessment framework?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Framework for a learning domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Does the framework include an explanation of the purpose of the assessment and who is being assessed?</td>
</tr>
<tr>
<td>4. Does the framework include a definition and description of the domains included in assessment?</td>
</tr>
<tr>
<td>5. Does the framework include an outline of the content areas of the domain that are included in the assessment?</td>
</tr>
<tr>
<td>6. Does the framework include an outline of the skills that are intended to be measured in the assessment?</td>
</tr>
<tr>
<td>7. Does the framework include specifications for the design and for operationalising the assessment (ie the assessment blueprint)?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contextual framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. Has there been research into what contextual influences would be expected to have an impact on the student learning outcomes in question for the study?</td>
</tr>
<tr>
<td>9. Have different levels of contextual factors been considered for the framework (eg the student level, the home environment, the school and classroom level, the wider community)?</td>
</tr>
<tr>
<td>10. Have the factors included in the contextual framework been well defined?</td>
</tr>
<tr>
<td>11. Is there a clear purpose for including the contextual factors selected?</td>
</tr>
<tr>
<td>12. Is relevant research cited, where available, to provide justification for including the contextual factors selected?</td>
</tr>
</tbody>
</table>
### Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blueprint</td>
<td>A description of how the test will be constructed, including the details of the proportion of items that will assess different domains and skills and the response formats.</td>
</tr>
<tr>
<td>Competency-based assessments</td>
<td>Assessments that are designed to measure generic knowledge and skills that are cross-curricular or that go beyond any particular school curriculum.</td>
</tr>
<tr>
<td>Constructed- response item</td>
<td>An item for which the student constructs, or generates, a response to the question.</td>
</tr>
<tr>
<td>Contextual information</td>
<td>Data collected through questionnaires on a range of topics that are useful to policy and in understanding the test results in context.</td>
</tr>
<tr>
<td>Curriculum-based assessments</td>
<td>Assessments that are based on measuring students’ understanding or knowledge of a curriculum or part of a curriculum.</td>
</tr>
<tr>
<td>Domain</td>
<td>The area of learning that is the focus of an assessment. This may be a curriculum area (e.g., mathematics or science), or more generic areas of learning (e.g., reading, writing or problem-solving).</td>
</tr>
<tr>
<td>Domain content areas</td>
<td>The content categories (or sub-domains) that are to be included in the test which are specific to the domain. For example, in mathematics, typical content categories are number, space, measurement and statistics.</td>
</tr>
<tr>
<td>Multiple-choice item</td>
<td>An item that presents several options as answers, from which the student selects one.</td>
</tr>
<tr>
<td>Response formats</td>
<td>The ways in which students need to respond to the items (e.g., multiple-choice, closed constructed-response, short open-response).</td>
</tr>
<tr>
<td>Skills</td>
<td>The ways of thinking, or intellectual approaches, that develop as individuals become increasingly proficient in a domain (sometimes called ‘processes’, ‘cognitive domains’ or ‘aspects’).</td>
</tr>
<tr>
<td>Target population</td>
<td>A particular group of people that the assessment is attempting to describe or measure outcomes for. For example, an assessment may aim to measure reading ability of Class III students in government schools in a particular state. This group of people is referred to as the target population.</td>
</tr>
<tr>
<td>Test developers</td>
<td>Those responsible for producing test content, including contributing to the test design, writing items, helping to interpret trial data to select items for the main survey, and interpreting main survey data for reporting.</td>
</tr>
</tbody>
</table>
References


7. ACER. (in press). SEA-PLM assessment framework. UNESCO.


Item writing
This chapter is intended to provide guidance for test developers who are creating items, that is, the questions or tasks for learning assessments.

The chapter describes the role of an assessment framework in relation to item writing; the characteristics and development of a good test instrument; guidelines for writing good items, including the construction of marking guides; and some of the essential factors that should be taken into account by test developers that relate to the design of a test. The characteristics of different response formats and the importance of a rigorous item review process at different stages of test development are covered. This chapter ends with a discussion of how establishing a team of experienced test developers is essential in creating a successful test program.

6.1 The importance of an assessment framework to item writing

A good test instrument depends upon a clear idea of why the test is developed, its objectives, and which elements of the domain (sometimes referred to as the ‘subject’ or ‘learning area’) are to be included in the assessment. An assessment framework provides this information, among other things, and is therefore very important for good item writing. Further information about the assessment framework can be found in Chapter 5.

The following parts of an assessment framework are particularly important for a test developer:

- test title and the target population
- fundamental purpose of the test
- the range of content and skills to be assessed
- the different response formats to be used
- the length of the assessment and the duration of the test.
6.2 Validity, reliability and fairness

The usefulness of an assessment instrument depends to a large extent on the comprehensiveness or coverage of the domains assessed (whether they be curriculum areas or more generic areas of learning), the consistency of marking (or scoring), and the reporting that allows accurate inferences to be made about what students know, understand and can do.

Validity, reliability and fairness are three important characteristics of assessment instruments. These three characteristics are briefly described below and further information is included in Chapter 4.

6.2.1 Validity

The content of the tests needs to be sufficiently representative of the domain that is being assessed. Tests should assess a range of levels of skills and knowledge. This should include skills at the target level, as well as skills that are at least one year above and two years below the target level. This allows the achievement of students at different levels of proficiency to be measured and described.

6.2.2 Reliability

The items that make up a test need to be designed and administered so that they are consistent in the scores they produce. Different versions of the same test should allow similar inferences to be drawn about the skills and knowledge of the students taking the tests. Where constructed-response items are included in a test (see Section 6.3.2), raters need to be consistent in the way they mark the items. Consistency in the measurement of the same population under different circumstances needs to be maintained: for this, adherence to an assessment framework and blueprint is required. Systems and support need to be in place to facilitate and standardise all aspects of administration.

6.2.3 Fairness

Test design and content must be fair to all students. The assessment must allow students to demonstrate their level of proficiency in the domain of interest, regardless of their background. It should be equitable to students of different genders, language backgrounds, cultural backgrounds and socio-economic backgrounds. For example, items about a popular male sport might disadvantage females.
6.3 The item writing process

This section describes the item writing process. In particular, it discusses finding or creating the stimulus material; the use of different response formats; and writing and reviewing the items.

6.3.1 Finding or producing the stimulus

‘Stimulus’ refers to the prompt or context on which an item is based. For example, in a reading test, the stimulus is often a prose text made up of one or more paragraphs. In a mathematics test, the stimulus may be a diagram or a graph.

Finding or creating a rich and interesting stimulus is a prerequisite for the item development process. The stimulus should not only be challenging, it should also offer opportunities to write good items. The stimulus should be attractive to students, be neither too hard nor too easy, and be accessible and equitable for different students in terms of content or language complexity, illustrations, and so on. Stimulus related to trauma, natural disasters, violence, bad language and contentious issues should be avoided. Importantly, the test developer must also verify the authenticity and any copyright issues of the stimulus before proceeding to item development.

Where a test is to be translated in more than one language, the stimulus should not pose any linguistic problems when translated. See Chapter 7 for a comprehensive discussion of issues related to the translation of instruments.

Often, more than one item is based on a single stimulus. Students only have to consider the stimulus once, and then they can use the information from it to answer several items. A ‘test unit’ is comprised of the stimulus material and the item(s) based on that stimulus. Exhibit 10 provides an example of a test unit which contains a stimulus (text and a graph) and two items (Questions 1 and 2). Students use the information provided in the graph to answer both of these items.

6.3.2 Using different response formats

The main response formats are multiple-choice items and constructed-response items. To answer a multiple-choice item, the student selects a response from provided options. To answer constructed-response items, the student constructs, or generates, his or her response. This section discusses multiple-choice items and two main types of constructed-response items: closed constructed-response and short open-response. Tests can also include essays or extended-response items. However, these are less commonly included in large-scale assessments.
6.3.2.1 Multiple-choice items

Multiple-choice items are written so that there is only one correct answer, so they can be reliably scored. Multiple-choice items are generally faster for students to answer compared to constructed-response items. Therefore, a greater number of items can be included in a test, meaning that a wide range of content can be covered. Multiple-choice items are also relatively easy to administer and mark, especially when marking is done by machines. For large samples, machine marking is quick and easy and saves costs. However, good multiple-choice items are not easy to construct and require skilled test developers. If distracters (incorrect responses) are not plausible (credible), it becomes easy for students to guess the correct response. This interferes with the accurate measurement of students' skills.

A multiple-choice item is made up of:

- a stem (the part of the item that introduces the options)
- one key (the correct or only acceptable answer)
- typically three or four distracters (incorrect options).

Items may refer to a stimulus material (see Chapter 8 for further details). The stimulus material provides information to answer the question and can include material such as text, graphs and diagrams. Exhibit 10 provides an example of a test unit made up of a stimulus and two items with the different components labelled.

**Exhibit 10:**
Example of a test unit with components labelled

![Graph showing speed of a racing car along a 3 km track](image)

**Question 1: SPEED OF RACING CAR**

What is the approximate distance from the starting line to the beginning of the longest straight section of the track?

- A 0.5 km
- B 1.5 km
- C 2.3 km
- D 2.6 km

**Question 2: SPEED OF RACING CAR**

Where was the lowest speed recorded during the second lap?

- A at the starting line.
- B at about 0.8 km.
- C at about 1.3 km.
- D halfway around the track.

Source: PISA Released Items 2006: Mathematics [1]
What makes a good multiple-choice item?
• It assesses a single well-defined competency.
• The stem is self-contained so that students understand clearly what is being asked of them.
• The stem contains as much of the item content as possible (see Exhibit 13).
• The stem and options are unambiguous and grammatically correct.
• The distracters are plausible but clearly wrong.
• There is only one key (correct or acceptable answer).
• The key does not stand out in any way: visually, grammatically or linguistically.
• The options are similar in structure, length and linguistic complexity.
• Options do not overlap in meaning.
• Negative wording is avoided.
• Words such as ‘always’, ‘never’ and ‘none of the above’ are avoided. The use of these terms can make it too easy for students to identify that these options are incorrect.

Exhibit 11, Exhibit 12 and Exhibit 13 illustrate some good and bad features of multiple-choice items.

**Exhibit 11:**
A good item: Multiple-choice item with plausible distracters

```
What is the area of the rectangle?

A. 3 m²
B. 16 m²
C. 32 m²
D. 48 m²
```

Source: International Benchmark Tests (IBT) sample test questions [2]

This is a strong item in that the distracters, options A, B and C, are all plausible and if selected would each indicate a common misconception. For example, a student who chose option C may think that the area of a rectangle is found by adding the four sides.

**Exhibit 12:**
A flawed item: Multiple-choice item with a stem that is not self-contained.

```
The red team and the blue team played a game. Here is the ball they played with.

The ball
A. is a cylinder.
B. is a sphere.
C. is a cube.
D. is a pyramid.
```
After reading the stem of a well-constructed item, the student will immediately understand what the item is asking. Even before reading the options the student will begin to predict what the correct response will be, and then will look for a match among the options. The student’s thinking will be activated. Exhibit 12 is flawed because after reading the stem (‘The ball’) the student would not know what was being asked. The student would need to read the options to find this out – a more passive approach, and an approach more likely to lead to guessing.

In Exhibit 13 the flaw has been removed: after reading the stem, the student will know what he or she needs to think about.

**Exhibit 13:**
A good item: Multiple-choice item with self-contained stem

The red team and the blue team played a game. Here is the ball they played with.

What shape is this ball?
A. It is a cylinder.
B. It is a sphere.
C. It is a cube.
D. It is a pyramid.

Source: MTEG Afghanistan Released Item [3]

### 6.3.2.2 Closed constructed-response items

Closed constructed-response items require a very short response such as one or two words, or a numeric response. They are called ‘closed’ because there is only one acceptable response, or a very restricted range of acceptable responses. Closed constructed-response is a useful format when the answer would be quite obvious in a multiple-choice format due to lack of plausible distracters. A disadvantage of closed constructed-response items is that they may need to be manually scored, which adds to the cost of the overall assessment process.

Even though the acceptable range of responses is very small, a marking guide must be carefully constructed, to allow for all possible varieties of acceptable response. What makes a good closed constructed-response item?

- The item is focused.
- Typically an answer of only one or two words, or a number, is required.
- A clear and comprehensive marking guide is developed.
Exhibit 14 is an example of a well-designed closed constructed-response item.

**Exhibit 14:**
A good item: A closed constructed-response item with a detailed marking guide

In a game, this 8-sided spinner is used. 
In the diagram the spinner shows ‘Miss a turn.’
The spinner was used 200 times.
About how many times would you expect the result ‘Go to jail’?

**Marking guide**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Code 1</td>
<td>Correct method but incorrect answer or incomplete</td>
</tr>
<tr>
<td>Code 0</td>
<td>Other</td>
</tr>
<tr>
<td>Code 9</td>
<td>Missing</td>
</tr>
</tbody>
</table>

**Source:** International Schools’ Assessment (ISA) sample questions [4]

Note that the marking guide for Exhibit 14 includes not just the most obvious answer (‘25’), but also alternatives that have been judged acceptable by the test developer (‘about 25’, ‘20–30’, etc.). Exhibit 14 also illustrates a ‘partial credit’ item: that is, a student can get full credit for the item (Code 2), partial credit for the item (Code 1), or no credit (Code 0). If a student has not responded, this is coded as missing (Code 9).13 In this case, full credit is obtained with the answer ‘25’ or equivalent; partial credit is obtained if the student shows his or her working, but does not provide the final (correct) calculation.

6.3.2.3 Short open-response items

Like closed constructed-response items, short open-response items require the student to generate or construct their answer, rather than select from a set of options. The distinguishing feature of open-response items (compared with closed constructed-response items) is that a wide range of responses is acceptable.

Responses to short open-response items include short written explanations, mathematical expressions, drawings and diagrams, analysis or evaluation. Adequate space on the paper and adequate time must be provided for answering such items.

A feature of open-response items is that they require high levels of item writing skill, including the construction of a marking guide comprehensive enough to accommodate a wide variety of answers, and clear enough to ensure consistency between different raters. Even with well-constructed marking guides, short open-response items are expensive and time-consuming to mark. Accordingly, to write this kind of item successfully the test developer must have a clear idea of what he or she intends to assess and be able to articulate rules for the range of responses that will be considered acceptable. It is also important that there is comprehensive training for raters.

---

13 A ‘9’ is the code that is commonly used to indicate missing data.
What makes a good short open-response item?

- The kind and extent of response expected is clear in the question.
- There are clear limitations on the kinds of responses that will be considered acceptable.
- The marking guide can be applied consistently and reliably, and accommodates a variety of responses.
- If there are different levels of credit (see Exhibit 14 above for an example), each level is clearly described.
- A range of example responses is provided for each level of credit to illustrate the marking rules.
- The marking guide focuses on the skills or knowledge that the item intends to measure, and not on skills irrelevant to that intent. (For example, if understanding of decimal fractions is being tested, poor spelling in the response is not penalised.)

Exhibit 15 shows an example of a well-constructed short open-response item from a reading test.

**Exhibit 15:**
A good item: Well-constructed short open-response item with a comprehensive marking guide

Happiness is different things to different people.

A sunny day makes me happy but farmers enjoy the rain.

Sara

Sara’s second sentence is an example that shows how ‘Happiness is different things to different people.’

Write another example that Sara could have used.
6.3.2.4 Summary of different response formats

Exhibit 16 provides a summary of some of the advantages and challenges associated with the three different response formats that have been discussed.

**Exhibit 16:**
Summary of the advantages and disadvantages of different response formats

<table>
<thead>
<tr>
<th>Response format</th>
<th>Advantages</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-choice</td>
<td>• Reliably scored</td>
<td>• Requires sufficient plausible distractors</td>
</tr>
<tr>
<td></td>
<td>• Relatively quick for students to answer, so many questions can be included in a test and therefore the test can cover a wide range of content</td>
<td>• Requires a high level of precision in development of stem and distracters</td>
</tr>
<tr>
<td></td>
<td>• Easy to administer and mark, therefore saving costs</td>
<td>• Does not promote generative or creative thinking</td>
</tr>
<tr>
<td>Closed constructed-response</td>
<td>• Can be used when there is a lack of plausible distracters for multiple-choice questions</td>
<td>• Requires manual scoring, which adds to costs</td>
</tr>
<tr>
<td></td>
<td>• Students can be credited for partial understanding (eg correct working but incorrect answer)</td>
<td>• Requires accurate and detailed marking guide development</td>
</tr>
<tr>
<td>Short open-response</td>
<td>• Can be used when there are a lack of plausible distracters for multiple-choice items</td>
<td>• Requires a high level of expertise to write both the items and the marking guide</td>
</tr>
<tr>
<td></td>
<td>• Students can be awarded for partial understanding</td>
<td>• Requires comprehensive training of raters</td>
</tr>
<tr>
<td></td>
<td>• Allows assessment of generative or creative thinking</td>
<td>• Expensive and time-consuming to mark</td>
</tr>
</tbody>
</table>

Source: International Schools’ Assessment (ISA) sample questions [4]
6.3.3 Initial drafting of items

Test developers design items according to the specifications in the framework and blueprint, including the response format, content and skills to be assessed. Test developers need to keep in mind the duration of the test and the wide range of abilities of the students in each class level.

6.3.3.1 Item descriptors

As well as thinking about the specifications of the framework, the test developer should create an item descriptor (sometimes called a question intent) at an early stage in the development of the item. The item descriptor describes what the item intends to measure. For example, the item descriptor for Exhibit 11 might be 'calculate the area of a rectangle using whole numbers of one and two digits'. The item descriptor for Exhibit 15 might be 'demonstrate understanding of an opinion by generating a different example expressing a similar opinion'. An item descriptor helps the test developer to focus on the purpose of the item, and how it relates to the domain. Later, it helps reviewers to evaluate whether the item's purpose is valid, and whether the item has fulfilled its purpose. Later still, item descriptors can be used in building reports on an assessment for teachers, parents, students and systems.

6.3.3.2 Intrinsic and extrinsic item difficulty

As mentioned above, items in an assessment should cover a wide range of difficulty, in order that the proficiency of students with varying levels of ability can be measured and described. It is important that even the students of lower ability levels can complete some of the items, so that the assessment shows what they can do as well as what they cannot do. It is equally important that the most able students are challenged by some items in the test. In both cases, and for all the levels of proficiency in between, the results will then indicate where it is appropriate to focus the next stages of teaching and learning.

Writing items with different intended levels of difficulty requires the test developer to have sound content knowledge and a good technical understanding of how items work. For example, in a reading assessment, items can be made more difficult by asking about information that is implicit rather than explicit in the text. In a mathematics assessment, all other things being equal, a problem that requires division is likely to be more difficult than a problem that requires multiplication. Clearly these examples focus on skills and knowledge that are intrinsic to the subjects of reading and mathematics respectively, and good test developers manipulate the levels of difficulty of items using these kinds of characteristics, which depend on the nature of the domain.

However, there are some factors that can contribute to item difficulty that are extrinsic to the domain or subject. These factors should not be used to make items more or less difficult because they do not contribute to outcomes that indicate students' skills and knowledge in the domain of interest.

Some of the features of items that make them unintentionally or extrinsically difficult are:

- long and complex sentences
- difficult vocabulary
- lack of clarity about the kind of response required (for constructed-response items)
- unnecessarily complex illustrations and diagrams
- a lot of reading (in a maths or science assessment)
- topics or texts that are not accessible to the students
- content or expression that favours a particular gender, group, community, religion or state.
6.3.4 Review process

The initial development of an item by an individual test developer is just the first step in the item writing process. In this section we discuss the steps in reviewing and further developing the items before they are ready to go into a live, ‘real’ test.

6.3.4.1 First review of items

Teamwork is an important aspect of the item writing process. In initial item development, a test developer may unintentionally show bias in development or choice of the stimulus, in the way the question or task is framed, or in the construction of the marking guide. A test developer may also unknowingly use ambiguous or confusing wording in the item. The review of items by peers to identify and rectify such problems at an early stage is an essential part of the item development process.

This initial review, sometimes called ‘panelling’ (or ‘item shredding’), is conducted by a group of peers. Test developers expert in the subject, including the test developer who initially developed the items, sit together to analyse sets of first-draft items. Items are inspected with regard to the:

- alignment with the assessment framework specifications
- significance of the skill assessed
- item difficulty (including distinguishing between intrinsic and extrinsic difficulty)
- precision and clarity of phrasing
- reasoning that gives the answer
- reliability and precision of the marking guide
- equity for all students of different genders, cultural backgrounds, languages, religion, state or socio-economic aspects.

Depending on the skills and experience of the item writing team, an item or set of items may go through a number of reviews and revisions of this kind before being prepared for a second major review, as described in the next section.

6.3.4.2 Second review following piloting

After the first round of revisions is completed, the items are tested on a group of students, perhaps in a single class (30 to 50 students). This is called a pilot test and is used to identify weaknesses or errors in the items. The piloting should be done with students who are as similar as possible to the students who will eventually take the test. If piloting involves a sufficiently large number of students, it can also help to establish the relative difficulty of the items. Students’ responses are carefully noted during piloting and the items are further revised based on the students’ feedback. Ideally, test developers conduct the pilots, so they can observe students’ responses first-hand, and talk to the students about what they are thinking as they do the tasks.
6.3.4.3 Third review following trialling

Once sets of items have been revised after piloting, they are tested on a larger scale in a field trial to assess their validity statistically. This is judged by whether: i) each item is contributing to measuring the domain of interest; that is, all the items ‘fit’; and ii) the items distinguish between students of different abilities (students who are more proficient in the domain or interest score more highly – get more of the items correct – than students who are less proficient). To collect evidence about these two features, each item needs to be trialled with at least 200 students. A statistical analysis is then performed to help identify items that best assess the designated skills. Information on item psychometric standards can be found in Chapter 4.

In a field trial, sometimes two different versions of an item are tried out with different groups of students, to see which version works better. The items are reviewed after the trial, and items that fail to perform well statistically are likely to be discarded. When reviewing performance of a set of items in a trial test, test developers need to take account of the framework and blueprint of the test alongside the item statistics. The final set of items should be statistically sound and also reflect the balance of content and difficulty levels as outlined in the blueprint.
6.4 Recording metadata

Item metadata is a record of all the information related to an item. Metadata can be kept in a spreadsheet or in a more sophisticated database.

A metadata record allows for systematic access to assessment material, which serves a number of purposes. An important purpose of recording the metadata for test developers is to assist in ensuring that items cover a range of content, skills, and levels of difficulty, as outlined in the framework.

Each item metadata record should include the following information:

- item code (unique item identifier) (see Section 8.1.5 for more information)
- the name of the test developer who wrote the item
- target population (e.g., Class III)
- domain/subject (e.g., mathematics)
- content classification (according to framework) (e.g., measurement)
- skill classification (according to framework) (e.g., applying)
- item descriptor (e.g., identifies the length of an object using a ruler)
- difficulty estimate (easy, medium, or difficult)
- marking guide or key (e.g., B)
- item number in test booklet (e.g., item 12).

Some of these pieces of metadata—content classification, skill classification, and item descriptor—should be developed at an early stage, at the same time as the item itself, to ensure that the item reflects essential elements of the framework. Other parts of the information cannot be recorded until a later stage of development. This includes the number of the item in a test booklet or form.
6.5 Training and building a team of test developers

As this chapter has shown, good item writing is a challenging task that requires specialised skills. Writing items based on scientific inquiry, real-life problem-solving, retrieving information or drawing inference from a text requires time and practice. For an institution (such as a ministry of education) that needs to deliver regular assessments, it is advisable to build up a team of dedicated test developers focused primarily on the task of developing the test instrument. Alternatively, the task of item development can be outsourced to an organisation that specialises in assessment. Many developed nations such as the United States, United Kingdom and Australia do practise selective outsourcing of assessments to credible third-party organisations that have a reputed track-record in assessment services.

Alternatively, a system may choose to employ teachers as test developers. Even for the most seasoned and experienced teachers, however, item writing skills will need to be learnt. A good teacher has excellent knowledge of his or her subject, but is unlikely to have acquired the technical knowledge of assessment that underpins strong test development. Organising workshops to develop teachers’ item writing skills is a useful way of training teachers.
6.6 Summary

If assessments are to provide accurate and useful information about student achievement, the quality of the test is of vital importance. As discussed in this chapter and in Chapter 5, the assessment framework forms the basis of the item development, including informing decisions about the response formats, range of item difficulties, domains and skills to be assessed.

Item writing is a complex process and it is important to have an experienced team of test developers who are familiar with the principles of good item writing and who also have domain expertise. Developing items is a resource-intensive process as there are multiple stages involved, including drafting, panelling, piloting, revising, field trialling and analysing items. These different stages help to ensure that the final test instruments are valid, reliable and fair.
### Checklist: Item writing

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Before item writing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Has an assessment framework been developed?</td>
<td></td>
</tr>
<tr>
<td>2. Does the test development team have experience with item development and domain expertise?</td>
<td></td>
</tr>
<tr>
<td><strong>Item writing</strong></td>
<td></td>
</tr>
<tr>
<td>3. Are the stimulus materials appropriate?</td>
<td></td>
</tr>
<tr>
<td>4. Are the item response formats suitable for the assessment?</td>
<td></td>
</tr>
<tr>
<td>5. Have the items been drafted following good item writing principles? (For example, for multiple-choice items is there only one correct answer and are the distracters plausible?)</td>
<td></td>
</tr>
<tr>
<td>6. Does each item have an item descriptor?</td>
<td></td>
</tr>
<tr>
<td>7. Do the items cover a range of difficulty levels?</td>
<td></td>
</tr>
<tr>
<td>8. Have the items been reviewed during a panelling process?</td>
<td></td>
</tr>
<tr>
<td><strong>Piloting and reviewing</strong></td>
<td></td>
</tr>
<tr>
<td>9. Has there been research into what contextual influences would be expected to have an impact on the student learning outcomes in question for the study?</td>
<td></td>
</tr>
<tr>
<td>10. Have different levels of contextual factors been considered for the framework (eg the student level, the home environment, the school and classroom level, the wider community)?</td>
<td></td>
</tr>
<tr>
<td><strong>Field trial and reviewing</strong></td>
<td></td>
</tr>
<tr>
<td>11. Have the items been field trialled?</td>
<td></td>
</tr>
<tr>
<td>12. Have the items been statistically analysed after the field trial?</td>
<td></td>
</tr>
<tr>
<td><strong>Items for main survey</strong></td>
<td></td>
</tr>
<tr>
<td>13. Are the final set of items statistically sound and do they also reflect the balance of content and difficulties as outlined in the blueprint?</td>
<td></td>
</tr>
<tr>
<td>14. Is the metadata for each item recorded?</td>
<td></td>
</tr>
</tbody>
</table>
## Blueprint
A description of how the test will be constructed, including the details of the proportion of items that will assess different domains and skills and the response formats.

## Constructed-response item
An item for which the student constructs, or generates, a response to the question.

## Distracters
The incorrect options provided in a multiple-choice item.

## Domain
The area of learning that is the focus of an assessment. This may be a curriculum area (e.g., mathematics or science), or more generic areas of learning (e.g., reading, writing or problem-solving).

## Field trial
Administration of items under test conditions, used to test the items' validity and the administration procedures. Occurs before the main survey with a sample of at least 200 students per item who are similar to the target population.

## Full credit
The maximum score for an item.

## Item descriptors
Description of what the item intends to measure.

## Item metadata
All the information related to an item (including the item code, target population, content and skill being assessed)

## Item statistics
The data used to assess whether items are functioning as they should (e.g., the data used to check whether items in different languages are equivalent).

## Items
The questions or tasks used in an assessment.

## Key
The correct or only acceptable answer to a multiple-choice item.

## Marking guide
The description of the marking/scoring categories that are used to categorise and score a student’s answer.

## Metadata
A record of all the information related to an item, including the item code, the domain and skills the item is assessing, the estimated difficulty level and the item descriptor.

## Multiple-choice item
An item that presents several options as answers, from which the student selects one.

## Panelling
A process where a group of test developers (including the test developers who drafted the items) review and evaluate the draft items, looking for ways to make improvements.
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial credit</td>
<td>Part of the available score for an item/task (e.g., one point out of a possible three).</td>
</tr>
<tr>
<td>Pilot test</td>
<td>A trial test involving a small sample of students.</td>
</tr>
<tr>
<td>Plausible distracters</td>
<td>Options provided in an item that are incorrect but credible (i.e., they are not the key).</td>
</tr>
<tr>
<td>Response formats</td>
<td>The ways in which students need to respond to the items (e.g., multiple-choice, closed constructed-response, short open-response).</td>
</tr>
<tr>
<td>Raters</td>
<td>Raters, or scorers, are the people responsible for marking the items or tasks.</td>
</tr>
<tr>
<td>Stem</td>
<td>The part of the item that contains the question or task (e.g., in a multiple-choice item, the part that introduces the options).</td>
</tr>
<tr>
<td>Stimulus material</td>
<td>The prompt or context on which one or more items is based. For example, in a reading test, the stimulus is often a prose text made up of one or more paragraphs. In a mathematics test, the stimulus may be a diagram or a graph.</td>
</tr>
<tr>
<td>Test developers</td>
<td>Those responsible for producing test content, including contributing to the test design, writing items, helping to interpret trial data to select items for the main survey, and interpreting main survey data for reporting.</td>
</tr>
<tr>
<td>Test unit</td>
<td>A stimulus material and the item(s) that are based on the stimulus.</td>
</tr>
</tbody>
</table>
References


Equivalence in multilingual assessments
Equivalence in multilingual assessments

It is important to ensure that students undertaking the same tests in different languages are responding to items and tests that are equivalent.

Items in different languages should measure the same knowledge, skills and processes as outlined in the assessment framework. (See Chapter 5 for more information on the role of the assessment framework.) The translation and quality assurance processes must be robust if the results from two or more language versions of a test are to be compared. This chapter describes how to assure translation quality from the initial stages of item writing. It then outlines some different methods and models to help ensure the quality of the translations. Lastly, it briefly discusses the statistical analyses needed to assess whether items in different languages are equivalent.

Throughout this chapter the following definitions are used:

- **Source:** The source describes the version of the units upon which all translations in the assessment are based. *Source language* is the language used in the source.
- **Target:** The target describes the translated version of the unit. *Target language* is the language used in the target.
7.1 Translation quality assurance

As noted above, items in different languages should measure the same knowledge, skills and processes. Translation quality assurance processes are necessary to ensure that tests in different languages are linguistically equivalent. Test equivalence means that students of the same ability level who are given tests in different languages should have an equal probability of answering an item correctly [1]. In other words, the language of the test should not affect the difficulty of an item or test.

Of course the need to maintain the meaning and difficulty of the items across all languages should be balanced with the need for the translations to be culturally appropriate. Balancing these requirements is a particular challenge for translators.

Test equivalence is especially important in India, where there are a vast number of languages spoken in the country and there is often a need to make valid comparisons of educational achievement across states, districts, schools and social groups. Some of the challenges particular to India highlight the need for robust translation designs and procedures. These challenges include:

- the diversity of language families
- the abundance of dialects within standard regional languages
- the varied use of gender across regional languages
- the variety of grammatical and punctuation usages across regional languages
- the common use of English words in everyday language and the decision as to when the use of these words in an assessment is appropriate [2].

Establishing test equivalence starts well before translation, at the very beginning stages of item writing. In the highest quality multilingual tests, translation issues are first considered as the item is being written. Test developers need to ensure that before beginning translation:

- the content of the stimulus material and items is finalised
- the style and register are appropriate
- the grammar and spelling are accurate
- the text has been checked and edited for translatability (see Section 7.1.1)
- item-specific translation guidelines are included where helpful (see Section 7.1.3).

7.1.1 Translatability

A translatability evaluation involves a panel of linguists undertaking a very quick and unpolished translation of the assessment stimulus and items. It is a cost-effective way to avoid future problems caused by items that are written in such a way as to be difficult to translate and adapt. From this activity, translation issues are identified and categorised. With this information, a senior linguist is able to produce a translatability report which includes general suggestions for improving the source version’s translatability (see Section 7.1.2) as well as item-specific suggestions (see Section 7.1.3). Test developers can use this report to edit the source units to improve translatability before sending them for translation.

If problems are identified and rectified early in the translation workflow, this will avoid potentially time-consuming and error-prone fixes later on in the process.

14There can also be a variety of usages within different schools and in different curriculum areas.
7.1.2 Generic translation guidelines

Before translation begins, a set of translation guidelines should be developed. This document should provide advice to translators on managing common translation issues in educational assessments. With the same advice provided to all translators, this helps ensure the consistency and efficiency of translations.

Translation guidelines can include rules for:

• adapting fictional and non-fictional names
• adapting mathematical symbols, place value and expressions
• adapting scientific terms (e.g., English, Latin)
• maintaining register and difficulty of vocabulary
• maintaining wording patterns in the text
• dealing with idiomatic expressions.

7.1.3 Item-specific translation guidelines

Along with the generic translation guidelines, item-specific translation guidelines should also be provided. The purpose of item-specific guidelines is to highlight exact text elements and provide specific advice for translating that text. Again, this helps to keep the translations consistent and avoids misinterpretations. Item-specific translation guidelines should be drafted by the test developers at the time of item writing.

Item-specific translation guidelines usually provide advice on:

• maintaining a pattern of synonyms or exact word matches
• clarifying the meaning of a word where there is a chance it might be misinterpreted
• dealing with ambiguous phrases or idioms
• dealing with situations where an adaptation should or should not be made, for example, using a local name, or maintaining foreign currency.

7.1.4 Finalisation of source versions

Before translation, source versions of the assessment material need to be finalised. It is always possible that an issue detected during translation could force a change in the source version. However, it is important to finalise the material as much as possible prior to the translation. This includes checking to ensure that the content, style and grammar are correct and final. It is important to finalise the source version to reduce problems with version control which can be caused by miscommunication about updates to units and document management.
7.1.5 Field trialling

All language versions of the test and questionnaire should be field trialled in order to develop a high-quality assessment. The sample size for the field trial must be large enough to provide stable statistical information. The sample size necessary will depend on the method chosen to analyse the data. (See Chapter 9 for more information about sampling.) Generally, 200 students per item is considered appropriate to achieve reliable item statistics. The layout and structure of the booklets should be the same across all language versions; this is important for both the field trial and for the main survey. For more information about test and questionnaire design, see Chapter 8.

It is strongly recommended that a field trial is conducted. If this is not possible, statistical analyses could still be carried out on the final dataset to check for equivalence (see Section 7.3). This introduces the risk, however, that some (or many) items could be found to be incomparable. This could invalidate any links made between language versions, meaning the results could not be compared.
7.2 Linguistic quality assurance

Linguistic quality assurance is an essential element in achieving the equivalence of two language versions of a test. Two forms of linguistic quality control that are commonly used are back translation and forward translation (also known as team translation [3] and verification [4]). Back-translation has been commonly used in the past; however, forward translation is today considered best practice for multilingual assessments.

7.2.1 Methods: Back translation

In back translation, the source version is translated into a target language by a person or a group of people. The quality of the translation is then checked by a different person or group of people by translating the target version back into the source language. Equivalence is then judged by comparing the source version with the back-translated version.

7.2.1.1 Advantages of back translation

The primary advantage of this design is that it allows the test developers and evaluation and assessment experts themselves to form a judgement about the quality of the translation and adaptation [5]. This is not possible in the forward translation design unless the test developers and evaluation and assessment experts are proficient in both the source and target languages.

Other advantages are that back translation can identify some of the problems caused with poor translations or adaptations. For example, if the meaning of a word has been misinterpreted, this will become evident when it is translated back to the source language.

7.2.1.2 Disadvantages of back translation

While being an advantage for test developers and evaluation and assessment experts, the fact that judgements are made in the source language rather than the target language, is also the primary disadvantage of back translation. While back translation may identify problems in word meaning, there are other significant issues that back translation will fail to identify. For example, if the target version copied the grammatical structure of the source version, this may make the target version unreadable, but it will not appear as such in the back-translated version.

Some other disadvantages are that back translation:

- may not expose problems with fluency or appropriateness of vocabulary or register
- may not reveal if item-specific translation notes have been followed
- is a time-consuming process with judgements being made about equivalence only after the items have been translated and then re-translated.
7.2.2 Methods: Forward translation

In forward translation the source version is translated by a person or a group of people. Another translator or group of translators then judges the quality of the translation by comparing the equivalence of the source and target versions. Forward translation methods are used in most high-profile international multilingual assessments including the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS) and the International Computer and Information Literacy study (ICILS).

7.2.2.1 Advantages of forward translation

The primary advantage of this design is that judgements about the equivalence of the target version to the source version are made directly in the languages involved. For example, this method can identify not only words with incorrect meaning, but also a failure to adapt the translation to the context, or the use of an inappropriate register.

This design can incorporate expertise from other fields to ensure that the translation is contextually appropriate. These experts could include target-language speaking experts in the subject being assessed, teachers, or even students.

7.2.2.2 Disadvantages of forward translation

The primary disadvantage of this design is that it relies on professional translators, rather than test developers or evaluation and assessment experts, to make judgements about the equivalence of the target version compared to the source version. Translators, for example, may not identify when the translation causes the language in the target version to be more difficult than the language in the source version. Or translators may not be familiar enough with the school context to judge whether certain adaptations are appropriate. While other experts may be involved in developing the translation, it is the translator’s role to verify and judge the equivalence of the final translated version.

7.2.3 Methods: International best practice

While each quality control method has its strengths, forward translation is superior to back translation and is considered best practice for assuring assessment equivalence [6]. Both back translation and forward translation pick up the accuracy of the meaning of the translation. However, forward translation has an advantage over back translation in that it also picks up the fluency and equivalence of register, and checks that all translation notes have been followed. Forward translation allows all errors or inappropriately translated elements to be identified and recorded in the target version with all suggestions and decisions documented for future reference.

However, neither a forward translation nor a back translation quality control design is sufficient to ensure the equivalence of two language versions of a test. Translators and reviewers are not able to identify all problems in the test items. As previously discussed, it is essential that all items are field trialled before they are used for comparative purposes. Field trialling provides data about the items which can be analysed to assess equivalence (see Section 7.3).
7.2.4 Models: Design considerations

Current best practice is to use the forward translation method and adapt it to design a translation quality control model to suit the specific needs of the assessment project. Some factors to consider are timelines, budgets, tools, available expertise and staff availability. All high-quality forward translation models, however, do have some factors in common. All high-quality forward translation models:

- are comprehensive (all items undergo the same treatment, not just a subset of items)
- involve multiple stages where each stage of the quality control is independent from the others
- record and document all edits, suggestions and decisions about the translations at every stage for future reference.

It is important that these factors are included in any model that is selected in order to ensure the high linguistic quality of the assessment.

7.2.4.1 Comprehensive application

The quality assurance model used should be comprehensive and applied to all items, not just to a subset of items. Each item presents its own challenges for translation and it is only through applying the model to the whole set of items that the quality can really be assured.

7.2.4.2 Multi-stage process

For the forward translation design to be effective, every stage of the quality control process should be independent of the other stages. This means that the people involved should be completely different, and that they should avoid consulting with each other during the translation or review process. Ideally, the model will include reviews by subject experts, evaluation and assessment experts, language experts, cultural advisors and gender sensitivity experts.

7.2.4.3 Documentation

Documentation of the decisions and suggestions made during translation quality control is important, not only for the immediate purpose of keeping a record of which changes were made and by whom, but for improving future translations. The documentation can be used to inform future general and item-specific translation guidelines and improve items.

The documentation is also an important reference for data analysts at the stage of data cleaning. If items behave in an unexpected way, data analysts can use the translation and adaptation documentation to investigate possible reasons behind the unexpected behaviour of the item.

Finally, it is important to maintain an archive of all final versions of the materials used in multilingual assessments. These materials may need to be consulted in the future, for example in a subsequent assessment cycle.
7.2.4.4 Other considerations

Decisions about whether to include other features in a translation quality assurance model will depend on the resources, timeline and context. The considerations listed in the remainder of this section are considered international best-practice, and should be seriously considered while designing the quality control model.

7.2.4.5 Translation staff and process

Best practice is that professional and formally qualified translators should be used to undertake the initial translation. Ideally, the professional translators will be familiar with the principles of good item writing as well as the construct (that is, the knowledge, skills or behaviours that the assessment measures). However, if translators do not have this knowledge, training should be provided before the translation task begins.

Using professional translators will help ensure the work is consistent as translators will be familiar with translation tools. Translation tools help the translator to build a database of translations of source text which they can use as reference when the same or similar text elements appear in a different context (called translation memory). This means that their work is not only more consistent, but also far more efficient.

In some cases it might be difficult to find professional translators at all, especially in less widely-spoken languages. If this is the case and non-professional translators need to be used, it is particularly important that a robust forward translation model is implemented.

7.2.4.6 Multiple source or reference languages

If the assessment is being translated into more than one language, an option to help ensure the quality of the translations is to use more than one source language, or to use one of the main target languages as a reference alongside the source.

If more than one source language is going to be used, it is imperative that the two sources be equivalent. Determining this is a time consuming and resource intensive process. It involves multiple reviews by subject experts and linguists of already verified translations.

If equivalence is not established, further translations could introduce even more linguistic inconsistencies as they are further removed from the original source version (ie they will be translations of a translation). (For a detailed description of the development of a second source version see Chapter 5 of the PISA 2012 technical report [4]).

An option which involves fewer resources is to use one of the main target languages as a reference alongside the source. The reference language version would need to undergo a rigorous translation quality assurance procedure to ensure it is consistent with the source language version. However, the reference language would not need to complete as many equivalence tests as a second source language requires (as it is used as a reference rather than a source). During the translation process, all languages should be translated from the same source version. The reference language should only be used to check for similarities and differences in meaning, style and expression.
7.2.5 Models: Examples of forward translation designs

It is often the time, professional resources and budget available that influence the design of a translation quality control model. There are many different examples of models used in multilingual assessments, but best practice models use the multi-stage model of independent translators and reviewers.

Three different multi-stage models are provided as examples of how the forward translation method can be customised to suit different project needs and constraints.

7.2.5.1 TRAPD, Janet Harkness

Professor Janet Harkness developed the Translation, Review, Adjudication, Pretest and Documentation (TRAPD) model [3]. This can be seen in Exhibit 17. As the name suggests, this is a forward translation or team translation design which involves different experts at different stages of the process. This example is not specific to any project and can be used as a template.

In this model, two versions of the source are translated by two independent translators. Both versions then go to a team of reviewers who merge the two versions together, discussing the merits of both and choosing the best elements of each. Subject and cultural experts would be best placed in this group. The reviewed translation is then checked and signed off during the adjudication process. The adjudicator, whose job it is to sign off on the translations, should be someone who has strong language ability in both the source and target language and they should be familiar with the assessment program. All items then go to a pretest or field trial, which may uncover problems in the translation that reviewers may have missed.

Throughout this process, all queries, suggestions, edits and decisions are documented using a customised tool which allows all involved to gain a better understanding of the final translation.

Exhibit 17:
TRAPD model
7.2.5.2 PISA 2012

PISA 2012 was adapted into 98 national versions of the assessment. The model which uses double translation and checks by various experts can be seen in Exhibit 18.

Translations were based on two different source language versions. Two translated versions were reconciled into a single version that incorporated the best of both versions. This version was then verified by a professional translator trained in assessment-specific issues. An expert in the construct being assessed reviewed the verified versions and highlighted any problems that needed to be corrected. The national team was given a chance to review the expert’s comments. Finally, all problems that the expert indicated were essential to fix were checked by another translator.

This is a comprehensive model involving the following roles: professional translators, language experts (reconciler and verifier), subject experts (referee) and cultural advisors (national reviewer).

For a detailed description of the translation and verification processes used in PISA, see Chapter 5 of the PISA 2012 technical report [4].

Exhibit 18:
PISA 2012 translation model
7.2.5.3 National Achievement Survey (NAS) model for Indo-Aryan languages

The NAS model which is shown in Exhibit 19 was used for the Class X assessment. This is an example of the implementation of a translation quality control model under very tight constraints. Access to professional translators was not available and time was limited.

The English source was translated by a group of subject experts who were ideally very experienced in item writing and review. As there was no access to professional translators, the accuracy of the translation was assisted with the availability of a Hindi reference version. The Hindi reference version completed a separate linguistic quality control process which involved the source being translated by two independent groups of subject experts and reconciled into one premium version by a third group of subject experts. However, this process was not extensive enough to guarantee equivalence of the Hindi version with the source. For this reason, the Hindi version was used for reference only, and not as a second source version.

Review of the translated items in Indo-Aryan languages was conducted by a second group of subject experts who verified the translations against the generic and item-specific guidelines. Finally, the reviewed version was sent to the state board coordinator, who accepted, implemented or rejected the second group’s suggestions and changes. Throughout the process, all decisions were documented.

**Exhibit 19:**
NAS translation model for Indo-Aryan languages
7.3 Statistical analysis of linguistic equivalence

Even if a robust linguistic quality assurance process is carried out according to international best practice, the equivalence of the two language versions cannot be confirmed until the item statistics are examined. While linguistic quality assurance is necessary for equivalence, it is not a sufficient condition.

Therefore, linguistic equivalence needs to be confirmed through statistical analyses. For example, there is an assumption that when the source language test is translated into multiple test languages, these translated tests are psychometrically independent of the language; for example, that speakers of the source language have the same test experience as speakers of the target languages. That is, language group is sometimes wrongly assumed to be a construct-irrelevant factor in the functioning of these tests. The item statistics needs to be analysed to see if the assumption of linguistic equivalence holds true. If the statistical analyses of the data indicate that there is not linguistic equivalence, then groups of students completing different language versions of the test cannot be compared.

There are a number of statistical procedures that can be used to check if the assumption of language group independence holds. The details of these procedures are beyond the scope of this chapter so only the overview of two well-established analytical methods will be presented here.

7.3.1 Differential item functioning

Differential item functioning (DIF) estimates the degree to which items behave differently across different groups when administered to students of the same ability. The idea is that when members of two language groups have the same ability, they should perform equally on a given item. If they do not perform equally between language groups, the items will need to be examined for any linguistic inconsistencies. For example, the following translation issues may have occurred:

- more content information was provided in the translation
- there was a mistranslation
- an aspect of the translation, not present in the source version, directed the student to a particular response option.

Translation is not the only cause of DIF among students who sit different language versions of the test. There may be other explanations for DIF. For example, perhaps the subject matter of the item is not covered in the curriculum followed by one language group. If DIF is present for an item, it is important to determine the reason behind the DIF as this will guide your decision about whether an item should be included or excluded from the main survey, and whether item modifications are necessary.
DIF analyses should be conducted after both the field trial and the main survey. After the field trial, if there are items containing language DIF due to the translation, there are a number of choices. These choices include to:

- **Omit the item from the main survey.** Single items or whole units can be removed from the item pool and not administered in the main survey. In order to be able to do this, more items will need to be developed for the field trial than will be needed in the main survey.
- **Rectify the translation problem and include it in the main survey.** If you are confident of the exact cause of the DIF between languages, you can fix the problem and still use the item in the main survey. This option does carry a risk however, as the item is essentially new and has not undergone field trialling. It might be the case that the fix introduces a different error.
- **Include the item without changes in the main survey.** If you include an item with an unacceptable level of language DIF in the main survey, the results can still be used for reporting on each population separately. You cannot, however, use the results to make comparisons between the two populations.

If, after the main survey, there are items containing language DIF and investigation has revealed that the DIF is due to the translation, the item should be removed from any cross-language comparison analyses. Further information about DIF can be found in Chapter 4.

### 7.3.2 Multi-group analysis of measurement invariance

A DIF analysis, which tests whether an item functions differently across groups, may not be technically applicable within the context of comparing group differences on translated items. This is because when the items are translated, there are inherent differences between each translated version and the groups being compared are therefore taking items that are not exactly the same.

The second, more statistically rigorous, procedure is to conduct a test of measurement invariance where the latent construct can be modelled to have indicator variables (the test items or test subscales) that are grouped based on language, and differences are tested through multiple group confirmatory factor analysis [see 7, 8, 9]. This procedure takes into account the relationship between the (observed) indicators and the latent construct, thus providing a more complete picture of the differences due to language groups.
7.4 Summary

In order for valid comparisons to be made between language groups, it is important that tests in different languages are equivalent. Translation issues should be considered and identified as early as possible at the time of item writing.

Once test materials have been written, a robust forward translation design should be applied. The design of the translation model will vary for different assessment programs and will depend on the resources available. However, all robust forward translation models should be applied to all items, involve multiple independent stages and should have a thorough documentation process. It is also essential that all test items are field trialled as the equivalence of the items in different languages can only be confirmed once the item data have been analysed.
### Checklist: Equivalence in multilingual assessments

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Planning for the translation process</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have sufficient resources been allocated to the translation (including budget, personnel and translation tools)?</td>
<td></td>
</tr>
<tr>
<td>2. Do program timelines allow for high-quality translation and quality control processes to be implemented?</td>
<td></td>
</tr>
<tr>
<td>3. Have different translation methods and models been considered and evaluated for their robustness and appropriateness?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>During item development</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Have item writers been given clear guidelines on translatability?</td>
<td></td>
</tr>
<tr>
<td>5. Have translatability evaluations been conducted (where appropriate)?</td>
<td></td>
</tr>
<tr>
<td>6. Have generic translation guidelines been developed?</td>
<td></td>
</tr>
<tr>
<td>7. Have item-specific translation guidelines been developed (where helpful)?</td>
<td></td>
</tr>
<tr>
<td>8. Have the source versions of the assessment material been finalised?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Translation and quality assurance processes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Have forward translation quality assurance methods been implemented?</td>
<td></td>
</tr>
<tr>
<td>10. Is the translation model comprehensive and applied to all items?</td>
<td></td>
</tr>
<tr>
<td>11. Has a multi-stage model of independent translators and reviewers been implemented?</td>
<td></td>
</tr>
<tr>
<td>12. Are decisions made during the translation process documented?</td>
<td></td>
</tr>
<tr>
<td>13. Are professional and formally qualified translators responsible for the translation (where possible)?</td>
<td></td>
</tr>
<tr>
<td>14. Does the translation process include reviews by subject experts, evaluation and assessment experts and other language and cultural experts?</td>
<td></td>
</tr>
<tr>
<td>15. If the assessment is being translated into multiple languages, have additional steps to ensure the translation quality been considered?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical analysis after the field trial</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16. Have all language versions of the test and questionnaires been field trialled?</td>
<td></td>
</tr>
<tr>
<td>17. Have DIF analyses and/or measurement invariance been conducted after the field trial?</td>
<td></td>
</tr>
<tr>
<td>18. Have decisions been made about items taking into account the field trial results? For example, attempting to fix the translation issue or dropping the item from the main survey pool.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical analysis after the main survey</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>19. Have DIF analyses and/or measurement invariance been conducted after the main survey?</td>
<td></td>
</tr>
<tr>
<td>20. Have items with problems due to translation been removed from cross-language comparison analyses?</td>
<td></td>
</tr>
<tr>
<td><strong>Glossary</strong></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td><strong>Adaptation</strong></td>
<td>Making changes to materials, such as items, in order to make them more suitable for the context.</td>
</tr>
<tr>
<td><strong>Adjudication</strong></td>
<td>The step in a translation process where the translation is signed off, indicating that it is ready for the next stage in the process (eg the field trial).</td>
</tr>
<tr>
<td><strong>Adjudicator</strong></td>
<td>The person responsible for signing off on the translation (see Adjudication).</td>
</tr>
<tr>
<td><strong>Construct</strong></td>
<td>What the assessment is intending to measure. This can include knowledge, skills and behaviours.</td>
</tr>
<tr>
<td><strong>Field trial</strong></td>
<td>Administration of items under test conditions, used to test the items’ validity and the administration procedures. Occurs before the main survey with a sample of at least 200 students per item who are similar to the target population.</td>
</tr>
<tr>
<td><strong>Generic translation guidelines</strong></td>
<td>Advice to translators on managing common translation issues in educational assessments.</td>
</tr>
<tr>
<td><strong>Idiomatic expressions</strong></td>
<td>Idiomatic expressions, or idioms, are phrases where the definition cannot be understood by the individual words (eg ‘two sides of the same coin’).</td>
</tr>
<tr>
<td><strong>Indicator variable</strong></td>
<td>Variables that can be observed and measured, which are deemed to represent a latent construct or factor (see Latent construct).</td>
</tr>
<tr>
<td><strong>Item statistics</strong></td>
<td>The data used to assess whether items are functioning as they should (eg the data used to check whether items in different languages are equivalent).</td>
</tr>
<tr>
<td><strong>Items</strong></td>
<td>The questions or tasks used in an assessment.</td>
</tr>
<tr>
<td><strong>Item-specific translation guidelines</strong></td>
<td>Advice to translators on translating specific parts of text.</td>
</tr>
<tr>
<td><strong>Latent construct</strong></td>
<td>Factors or traits that are not directly observable, such as maths ability.</td>
</tr>
<tr>
<td><strong>Measurement invariance</strong></td>
<td>The latent construct needs to be derived from a set of observed or indicator variables (see Indicator variable).</td>
</tr>
<tr>
<td><strong>Mistranslation</strong></td>
<td>A statistical property of a measure that relates to the extent to which the measurement of a construct is the same (invariant) regardless of the groups being measured.</td>
</tr>
<tr>
<td><strong>Multi-group confirmatory</strong></td>
<td>An incorrect translation.</td>
</tr>
<tr>
<td><strong>Psychometrics</strong></td>
<td>Theory and methods of measuring psychological traits, such as mathematical ability or motivation to read.</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Quality assurance</strong></td>
<td>Methods, procedures and checks used to ensure quality, in this case, the quality of the translations.</td>
</tr>
<tr>
<td><strong>Reconciliation</strong></td>
<td>Process of combining the best elements of multiple translations into one single version.</td>
</tr>
<tr>
<td><strong>Register</strong></td>
<td>The level of language used in a particular situation (for example, an informal register, a technical register).</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>The version of the units upon which all translations in the assessment are based.</td>
</tr>
<tr>
<td><strong>Source language</strong></td>
<td>The language used in the source version.</td>
</tr>
<tr>
<td><strong>Stimulus material</strong></td>
<td>The prompt or context on which one or more items is based. For example, in a reading test, the stimulus is often a prose text made up of one or more paragraphs. In a mathematics test, the stimulus may be a diagram or a graph.</td>
</tr>
<tr>
<td><strong>Target</strong></td>
<td>The translated version of the unit.</td>
</tr>
<tr>
<td><strong>Target language</strong></td>
<td>The language used in the target version.</td>
</tr>
<tr>
<td><strong>Test developers</strong></td>
<td>Those responsible for producing test content, including contributing to the test design, writing items, helping to interpret trial data to select items for the main survey, and interpreting main survey data for reporting.</td>
</tr>
<tr>
<td><strong>Test equivalence</strong></td>
<td>Students of the same ability level who are given tests in different languages have an equal probability of answering an item correctly.</td>
</tr>
<tr>
<td><strong>Translator</strong></td>
<td>Someone who translates text from one language into another language.</td>
</tr>
<tr>
<td><strong>Translatability evaluation</strong></td>
<td>A quick translation of the assessment stimulus and items to identify any potential translation issues.</td>
</tr>
<tr>
<td><strong>Unit</strong></td>
<td>A group of items that share a common stimulus material.</td>
</tr>
<tr>
<td><strong>Verification</strong></td>
<td>The checking of a translation to ensure that it is accurate.</td>
</tr>
<tr>
<td><strong>Version control</strong></td>
<td>Managing multiple versions of a document or file.</td>
</tr>
</tbody>
</table>


Test and questionnaire design
Test and questionnaire design

There are many decisions which need to be made when designing tests and questionnaires, including around the content, delivery, structure, analysis and consultation process.

Test and questionnaire design are essential steps in the assessment development process. The first part of this chapter (Section 8.1) outlines some of the important issues in test design. The second part of this chapter (Section 8.2) discusses questionnaire design\textsuperscript{15}. This chapter is linked with Chapter 5, which discusses the assessment framework and Chapter 6, on item writing. The topics of these three chapters together deal with the fundamental components of instrument development.

\textsuperscript{15}Although the topics discussed in this chapter are generally applicable to a wide range of assessment types, the scope of this chapter is mainly for large-scale assessments that have paper-based administration process.
8.1 Test design

In educational research, large-scale assessments provide an estimate of student achievement in one or several domains or subject areas. The design of appropriate assessment instruments is of crucial importance. A range of experts should be involved in designing the tests to ensure that data collection and reporting requirements are met, and that the implications for national policy are considered. The following sections look at the various processes involved in designing and implementing appropriate test instruments.

8.1.1 Test blueprint

The test blueprint (also sometimes called the ‘table of specifications’) is a very important document that determines how the test is constructed, and subsequently how the resulting test data are analysed and reported. The test blueprint is part of the assessment framework and has been described in Chapter 5. As has been discussed in Chapter 5, a test blueprint should specify things such as the domains to be assessed, the item response formats that will be used and the test length. The test blueprint provides guidance for the item writing (see Chapter 6) and test design.

8.1.2 Test length

Longer test booklets have the advantage of being able to include a wide range of interesting and engaging items and stimulus materials. However, shorter test booklets are easier and less expensive to produce, administer and mark, and are less likely to tire students. There are several factors that determine test length [1]:

- the number of domains to be covered
- breadth of coverage within domains
- item response format
- use of illustrations
- length of stimulus materials
- font size
- funding for printing.

Before item writing begins, a decision should be made about how many pages will be included in the final test(s). Item layout should always be clear and uncluttered; therefore, test developers will need to modify stimulus texts and illustrations if the test length is going to be limited.

It is also important to consider the amount of time students will take to complete the test. Several factors can affect the time students take to complete a test, such as the age of the students (older students can handle longer test times) and students’ familiarity with the item response formats in the assessment, as well as the difficulty of the items included in the test. Enough time should be allowed to enable most students to attempt most items. The longer the test, the more time needs to be allowed. It is recommended that information about the amount of time taken by students to complete test items is collected during piloting or field trialling.
8.1.3 Linking design

In large-scale assessments, each domain of interest includes a wide range of content and skills to be assessed. However, it is important that the test is not too long as this could overburden the students. It may be necessary to design the assessment so that each student is only administered a fraction of all available items in the assessment. This helps to ensure that the assessment sufficiently covers the content, while not overburdening students. In this case, the test developer needs to create a series of linked test booklets that are carefully designed and administered so that the data collected from them can be combined onto a single scale.

In general, administering different tests measuring the same construct (e.g., mathematics) to different groups of students results in separate scales. Even though the tests measure the same underlying ability (mathematics), the results are not comparable unless the different scales are linked, or merged, into one. To link one set of items to another the tests need to i) measure the same construct (e.g., mathematics) and ii) have either some items in common or some students in common.

With common item linking, two tests that measure the same construct will be administered to different groups of students. To link these two tests, there needs to be a subset of items that are common to both tests. These are referred to as link or anchor items. An example of common item linking is shown in Exhibit 20. As can be seen in Exhibit 20, both Test 1 and Test 2 contain a common subset of items – Cluster B.

**Exhibit 20:**
Two tests showing common item linking

<table>
<thead>
<tr>
<th></th>
<th>Cluster A</th>
<th>Cluster B</th>
<th>Cluster C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test 1</strong></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td><strong>Test 2</strong></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

With common person linking, two tests that measure the same ability but have no items in common are administered to the same subgroup of students. In both common item linking and common person linking, statistical methods are used to link or equate the two tests so that they are on the same scale. Chapter 12 provides further information about these two designs, in addition to information about other linking designs.

In the case where there are a large number of items distributed between multiple booklets, linking can also be used to:

- estimate change over time (e.g., comparing Class III performance in 2012 to Class V performance in 2015)
- estimate growth between class (grade) levels (e.g., comparing Class III performance in 2015 to Class V performance in 2015).

These two types of linking are known as horizontal and vertical linking, respectively.
Where there are multiple test booklets measuring the same construct, it is best practice to use common item linking to link the tests. This means that the results can be reported on one scale. However, when using this linking method it is important to select link items that display good statistical characteristics and cover a wide range of difficulties.

It is important that the link items function in the same way across the different tests. If the link items appear differently in different tests, this affects how they function. If their functioning is affected too much, the linking of the tests is weakened and the quality of the overall analysis is consequently undermined. Factors which affect how link items function are:

- wording of test item or stimulus
- wording and ordering of response categories for multiple-choice items
- item layout
- location of item within test
- context of the item (following easy or hard questions)
- changes in the marking guide for constructed-response questions
- changes in the curriculum/teaching
- the total set of link items needs to be representative of the whole test in:
  - average difficulty
  - item response format
  - subscale (if applicable).

It is also very important that link items perform similarly in different groups of students. They should not favour one group of students over another (e.g., boys compared to girls).

The number of link items required for a strong statistical link can vary, but normally between 8 and 10 common items are required [1]. Link items should be placed towards the beginning or middle of the tests so that all students have the opportunity to attempt them. They should also appear in a similar order in each of the booklets they appear in so that any differences in student performance is not caused by a change in position or ordering.

Link items should be of average difficulty, so that students of an average ability should have between 40% and 60% chance of answering the items correctly [1]. During piloting or field trialling, it is better to have many more link items, since the performance of the items is unknown. Following analysis of the field trial results, the best performing link items can be selected for the final test forms. Data analysts should always be involved in the process of selecting the link items.

8.1.4 Cluster booklet rotation design

Multiple test booklets are used to ensure that there is sufficient content coverage while keeping the testing burden on individual students low. These booklets then need to be linked using statistical methods so that the data collected from them can be combined onto a single scale. Common item linking is straightforward in cases where there are just two booklets, but what happens when there are more than two booklets? As with the earlier example, the booklets need to be carefully designed so that they can be linked with the appropriate statistical methods.

Many of the large-scale international assessments – such as the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS) and the Programme for International Student Assessment (PISA) – deal with this issue by creating mutually exclusive clusters of test items, then assigning those clusters
to booklets in a systematic fashion. An item cluster is a small group of test items or test units that are grouped together and treated as a block during test construction. In such a design, each booklet is made up of the same number of clusters (e.g., four clusters), and each cluster appears in more than one test booklet, as well as in a different position in each booklet. By allocating item clusters to more than one booklet, this ensures that the scores from different booklets can be linked and related to a common scale. Additionally, having each cluster appear in a different position in each booklet minimises position effects, whereby questions at the end of booklets appear harder as not all students are able to complete them.

The test design shown in Exhibit 21 is an example of a rotated test design from PISA. This particular design is known as a balanced incomplete block design. Test items were allocated to 13 test clusters (seven science clusters S1 to S7, two reading clusters R1 and R2, and four mathematics clusters M1 to M4). Each cluster appears in each of the four possible positions within a booklet once. An additional feature of this design is that each possible pair of clusters appears in one (and only one) booklet.

**Exhibit 21:**
Cluster rotation design used to form test booklets for PISA 2006 [2]

<table>
<thead>
<tr>
<th>Booklet</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S1</td>
<td>S2</td>
<td>S4</td>
<td>S7</td>
</tr>
<tr>
<td>2</td>
<td>S2</td>
<td>S3</td>
<td>M3</td>
<td>R1</td>
</tr>
<tr>
<td>3</td>
<td>S3</td>
<td>S4</td>
<td>M4</td>
<td>M1</td>
</tr>
<tr>
<td>4</td>
<td>S4</td>
<td>M3</td>
<td>S5</td>
<td>M2</td>
</tr>
<tr>
<td>5</td>
<td>S5</td>
<td>S6</td>
<td>S7</td>
<td>S3</td>
</tr>
<tr>
<td>6</td>
<td>S6</td>
<td>R2</td>
<td>R1</td>
<td>S4</td>
</tr>
<tr>
<td>7</td>
<td>S7</td>
<td>R1</td>
<td>M2</td>
<td>M4</td>
</tr>
<tr>
<td>8</td>
<td>M1</td>
<td>M2</td>
<td>S2</td>
<td>S6</td>
</tr>
<tr>
<td>9</td>
<td>M2</td>
<td>S1</td>
<td>S3</td>
<td>R2</td>
</tr>
<tr>
<td>10</td>
<td>M3</td>
<td>M4</td>
<td>S6</td>
<td>S1</td>
</tr>
<tr>
<td>11</td>
<td>M4</td>
<td>S5</td>
<td>R2</td>
<td>S2</td>
</tr>
<tr>
<td>12</td>
<td>R1</td>
<td>M1</td>
<td>S1</td>
<td>S5</td>
</tr>
<tr>
<td>13</td>
<td>R2</td>
<td>S7</td>
<td>M1</td>
<td>M3</td>
</tr>
</tbody>
</table>

When test items are arranged in clusters, it is best to link booklets with items taken from two or more clusters, just in case the items associated with one cluster do not perform as well. It is not necessary to use all items in a cluster for linking purposes; some items in a cluster can be unique.
8.1.5 Test structure and layout

The front page of the test booklet can be used for the collection of student demographic information, as well as information required for student tracking. (See Chapter 10 for more information about the student attendance sheet.). It should request information such as the student’s school, full name, class (grade) level, gender, language background (if necessary), and age.

It is also useful to include an option for the field investigator to record information such as whether students missed all or part of the test for any reason or whether any students required special assistance to help them complete the test. It is strongly advised that a data analyst is consulted about how this information is recorded and entered in data analysis software to ensure it is in a form that is appropriate for analysis.

The first few pages of the test booklet usually contain instructions for the students on how to complete the test, as well as some practice test items. Students must clearly understand how to respond to the test items. If separate answer sheets have been provided for students to record their answers to the test items, these answer sheets should clearly show how to link the item response or option printed on the test booklet with the appropriate position on the answer sheet. If the students are unfamiliar with any of the response formats included in the test, they should be given sufficient time and opportunity to attempt practice items before the formal test begins.

It is useful to have unique labels printed in greyscale next to the item. This is particularly useful when multiple test booklets are used. Items with the same label should be identical in every respect except for their order of appearance in a test booklet. A document should be produced for the data analyst that lists each item’s unique label. When multiple test booklets have been used, this document should include a list of which booklet(s) each item appears in and the order in which it appears.

How data will be collected from test booklets affects how each item should appear. Multiple-choice responses can be entered into data collection software manually or they can be electronically scanned. Electronic scanning of responses requires more specialised equipment and technical support, and the items’ responses need to be in a particular format (such as shaded circles) for the data to be collected. Manual data entry allows for a wider range of multiple-choice response formats (ticking boxes, drawing circles, underlining text), but using a consistent layout improves the efficiency of the data entry process.

It can also be useful to include the scoring options for certain items (eg closed constructed-response, or short open-response) in greyscale. Inclusion of these scores serve two purposes: they remind raters of the range of scoring options and raters can then simply circle the appropriate score for each item.

The layout of items should always ensure that the necessary amount of space has been allowed for i) the student’s response and ii) the rater’s score. An example can be seen in Exhibit 22 where the dotted line allows sufficient space for the student’s answer and the rater can indicate their score of 0, 1 or 9.16

16In this example a score of 0 indicates an incorrect answer, 1 indicates a correct answer and 9 indicates a missing response.
R040 LAKE CHAD

Figure 1 shows changing levels of Lake Chad, in Saharan North Africa. Lake Chad disappeared completely in about 20 000 BC, during the last Ice Age. In about 11 000 BC it reappeared. Today, its level is about the same as it was in AD 1000.

Question 3A: LAKE CHAD

In about which year does the graph in Figure 1 start?

Source: PISA Released Items 2006: Reading [3]
Proofreading is an extremely important process that should be performed before test booklets are sent for printing. It is especially important for piloting or field trialling. Test items and stimulus material (reading passages, graphics, illustrations) should appear in the field trial exactly as they will in the final test. Incorrect field trial items should not be used in the final booklets. Therefore, it is important to ensure that items in the field trial do not contain typographical errors and inconsistencies. It is crucial that all test booklets are thoroughly proofread. It is advisable that proofreaders attempt all items as though they were actually taking the test.

When reviewing test booklets, proofreaders should ensure that:

- initial instructions and practice items are clear and unambiguous
- items are clear and unambiguous
- stimulus material is clear and easy to read
- multiple-choice options include one correct answer with other options all clearly incorrect
- each of the multiple-choice options makes sense
- the items in a cluster are independent; that is, the answer to one item is not given in the stem or options of another item
- link items are identical
- no spelling or grammatical errors occur
- the layout of the various test booklets is consistent [1].

Booklets should also be rechecked when they have returned from the printer. It is important to check that all pages have been printed clearly, appear in the correct order, and are not duplicated. Illustrations should be clear and the stimulus material and corresponding items should be printed on the correct pages. Paper and printing quality should be sufficient so that items printed on one page do not interfere with the legibility of items on the surrounding pages. The test printing schedule should allow time for reprinting, if necessary, before the booklets are sent out to schools.

Test booklets should be structured so that they begin with some easy items to encourage lower-ability students to attempt the test. If a rotated cluster design has been used, it is beneficial to structure the clusters in this way as well so that it doesn’t matter which cluster is allocated to the first position in each booklet. It is then often desirable to mix the difficulty of the following items so that students do not abandon the test when they encounter a run of difficult items. It is also important to have some harder items earlier in the test so that students who work more slowly have an opportunity to attempt some harder items. As lower-ability students are less likely to finish the test, it is good to end with some harder items to challenge those students who can reach them.

When using multiple booklets, the overall difficulty of each booklet should be roughly the same. If any one booklet contains too many hard items, students may abandon the test, resulting in a lack of data for items appearing at the end of that test booklet.
8.2 Questionnaire design

Questionnaires are critical to large-scale survey research and provide complimentary data to achievement tests. While tests can demonstrate the achievement level of a student in relation to his or her peers, it is the content obtained from the questionnaire that puts the test results in context. Data from questionnaires can help illustrate factors associated with high-performing students, or perhaps more importantly, factors associated with low-performing students. Identifying which students are performing below a defined standard may be important to know, but understanding why they are performing at a lower level is arguably of greater importance.

8.2.1 Development process

8.2.1.1 Establishing content areas and questionnaire types

The first and most important part of questionnaire design is determining the areas to be explored. Ideally this has already been decided on as part of the contextual framework (see Chapter 5). The first question to be asked is: What is the purpose of the whole assessment? For example, a national or regional assessment designed to monitor standards in literacy and numeracy would likely want to investigate factors believed to influence these areas of learning. (See Chapter 2 for information about assessment purposes.)

Relevant stakeholders should be consulted to help determine what is to be examined through the questionnaire. This should include those who are funding or initiating the assessment program and other groups that may have an interest in the outcomes of the program, for example teacher unions where teacher data are being collected.

Once the content areas to be explored have been defined, the next step is determining the ways in which this data will be collected. In surveys assessing achievement in students, it is common to use a student questionnaire. The age of the students being assessed will need to be considered when deciding whether to use a student questionnaire; student questionnaires are rarely used with children under the age of eight.

Content areas that might be used in a student questionnaire include:

- respondent demographic characteristics
- social class, ethnic background and language spoken
- home environment
- home resources
- educational background
- mode of attendance for school
- opportunities and support for attending school
- health and nutrition
- attitudes to and perceptions of other students, school, teachers and homework
- behaviours and behavioural intentions.

A large volume of research emphasises the importance of classroom influences for student learning. Therefore, it is also common to implement a teacher questionnaire. In such instances, teachers are generally sampled from the same schools as students. During the sampling design stage, consideration needs to be given to whether teachers should be sampled, so that the classroom teachers of the students selected are all included.
Teacher questionnaires are also used in some large-scale research studies where teacher- and school-level information is the primary purpose of the study, and no student-level data are collected. The Teaching and Learning International Survey [4] and the Teacher Education and Study in Mathematics [5] studies are examples of larger studies of this type. Teacher questionnaires may be for all teachers in a school, all teachers of a particular class (grade) or all teachers of a specific subject (e.g., a numeracy study may only have interest in questioning mathematics teachers).

Content areas for a teacher questionnaire may include areas such as:

- respondent demographic characteristics
- educational background, professional development, training background and experience
- classroom and school resource availability
- attitudes, perceptions and satisfaction levels at school
- relationships with other teachers, the school head, parents, students and the community
- teaching methodology and style.

Most studies tend to collect data at the school level, through a school questionnaire. These are most often completed by the school head or a delegate. For assessments focused on specific areas, responsibility for parts of the questionnaire might be directed towards a specific role within the school. For example, a study on computers may require input from the school information technology coordinator, or a study on libraries would likely involve input from a librarian. Content areas typically found in school questionnaires include:

- respondent demographic characteristics
- educational background and experience
- school size
- school characteristics and policies
- school and community resources
- teacher education and professional development practices
- relationships with students, teachers, parents and the community.

8.2.1.2 Setting up the processes

Once the decisions have been made about the content areas and questionnaire types, the next step is to develop the questions themselves for the questionnaires. A suitable questionnaire development team should be set up with people who are knowledgeable about the purpose of the questionnaires. It is recommended that someone with expertise in data analysis is included in this team. The data analyst would be expected to have a perspective on the types of question formats that are more likely to produce more valid data, and may help avoid question structure issues that may result in unusable data for analyses.

All questions to be included should be part of a well-developed process. It is important to allow enough time for questions to go through a thorough review and editing process where other team members have time to make appropriate changes. It is helpful for each member of the group developing the questionnaires to complete the questionnaire as if they were a respondent. Any instance in which there is uncertainty or language which they think can be improved should be noted and addressed in revisions of the questions. The refinement will ideally take place over an extended period of time to allow several revisions of the questionnaire. Once the team is satisfied with the questionnaire, it should be distributed to the relevant stakeholders in the project for their feedback.
It is also suggested that the development team read through the questionnaire content of other large-scale surveys, particularly those from an Indian context. This can help to get an idea of the ways in which suitable questions can be asked. Two examples on the international scale were the participation of Himachal Pradesh and Tamil Nadu in the PISA project [6] and the Rajasthan and Orissa use of TIMSS 2003 material [7]. Established questions that have been shown to work well in other questionnaires could be useful also. Before questions are taken from other questionnaires, permission must be granted by the owners of that questionnaire material. It may be useful to adapt socio-economic status (SES) measures from other Indian studies that have successfully demonstrated the link between SES and student outcomes.

Once the questionnaire content has been finalised, it is strongly recommended to pilot the new material, whether part of a field trial or done on a smaller scale. This would involve similar procedures that are used for piloting test items. Ideally there would be a sufficient number of respondents complete the questionnaires to see whether the questions are working as intended. It would also be ideal to have qualitative feedback to allow respondents to give their thoughts on which questions were difficult or unclear. In each case, the respondents involved in piloting should be as similar as possible to those who will be targeted by the questionnaires in the main study.

8.2.2 Questionnaire item writing

8.2.2.1 Considerations for questionnaire item writing

When writing questions, it is important to try and be as clear and specific as possible, and to try to avoid potential ambiguity in understanding the question meaning. If the questionnaires are to be used across a variety of contexts, consideration should be given to the language and cultural factors.

When writing questions, an essential consideration should be the intention of using the data for that question. It can be highly tempting to come up with a long list of question areas or adapt questions from other questionnaires, in the hope that something interesting might come up at the analysis phase. However, this is potentially a waste of resources and not a suggested approach to developing questionnaire content. It is for this reason that stakeholders should be involved in determining what question areas are of interest.

Consideration needs to be given as to whether questions need to be asked, if reliable information is already available. For instance, a school questionnaire may ask about the type of school (whether it is public or private) and the size of the community the school is located in. This information might already be available to the team from the development of the sampling frame. There is no need to waste the time asking respondents to enter this information (and the resources associated with data entry) if this information is already known. An exception to this is if you want to confirm the data. Large-scale assessments often ask students to indicate their gender and age, even if they already have this information obtained from school records. Having this self-reported data can be used to confirm the identity of the student.

The questions need to be designed for the target population. Consideration needs to be given as to whether the respondents will be able or willing to give honest answers to questions. For example, respondents may not want to answer questions about personal income or about whether they have committed a crime.
Questionnaires developed for children, particularly those at a primary or early secondary level, need to be written in relatively simple language, and text kept as brief as possible without compromising the detail required for the questions. Although it would be expected that teachers and school heads would be able to handle a higher reading load, it is still encouraged to keep things as brief as possible so that the respondent remains engaged.

A related point is the length of the questionnaire. Longer questionnaires may put respondents off from completing them at all. Quite often teachers and school heads have very busy workloads and the length of the questionnaire might determine whether they choose to complete it or not. Longer questionnaires tend to produce more unreliable data as motivational factors influence responses. There tends to be more missing data towards the end of the questionnaire as respondents tire. As a general rule, no questionnaire completed by students should take longer than 20 minutes to complete (adults may be given questionnaires up to 30 minutes to complete).

8.2.2.2 Question formats

Questions can be asked in a range of different ways. The following section lists the major types of questions typically used in questionnaires from large-scale research surveys.

Open-ended questions require respondents to write an answer to a question as opposed to ticking a box or entering a number. The complexity of questions range from the simple (such as asking the respondent to write their name) to the complex (such as asking them to give a detailed list or reasoning). These types of questions can provide valuable qualitative data when used appropriately that may otherwise not be able to be obtained with other forms of questions. However, they are typically more demanding and time-consuming for the respondent. Therefore open-ended questions should only be included where specific information is wanted that is highly important and cannot be asked using other question forms.

Consideration needs to be given as to how the responses to open-ended questions will be analysed. If hundreds or thousands of questionnaires are completed, will someone code or analyse all of the responses? An example of an open-ended question is shown in Exhibit 23.

Exhibit 23:
Example of an open-ended question

What are your greatest needs in order to improve your teaching of science in your class?

Please write your answer on the lines provided below.
Categorical questions require respondents to pick from a list of categories which don’t have any order to them. Examples of categorical questions include gender, social class and building material type. Each category is assigned a number for data entry (for example a question on gender might use 1 for females and 2 for males). An example of a categorical question is shown in shown in Exhibit 24.

**Exhibit 24:**
Example of a categorical question

Which of the following is your favourite food?

*Please mark an ‘X’ in one square only.*

- Rice
- Roti
- Dal
- Samosa
- Another food

Questions using Likert scales allow researchers to quantify data that are typically qualitative in nature (e.g., levels of agreement, confidence and satisfaction). An example of a Likert scale is shown in Exhibit 25 below.

There is an inherent order to Likert scales and the number of options on a Likert scale typically ranges from as low as three options to a high of ten options. A Likert scale is a form of ordinal measurement, and like all ordinal scales the distance between points are not necessarily equivalent (e.g., the distance between ‘strongly agree’ and ‘agree’ is not necessarily equivalent to ‘agree’ and ‘disagree’). As is shown in this example, Likert scales do not always have a ‘neutral’ option. This can help avoid confusion around the term ‘neutral’ as this can be interpreted as neither agreeing or disagreeing, or may be interpreted as ‘I don’t know’.

**Exhibit 25:**
Example of a question with a Likert scale

To what extent do you agree or disagree with the following statements about your school?

*Please tick one box on each line.*

<table>
<thead>
<tr>
<th></th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) I like to read books during my lunch break</td>
<td>![1]</td>
<td>![2]</td>
<td>![3]</td>
<td>![4]</td>
</tr>
</tbody>
</table>
Other question formats include:

- **Frequency scales**, in which respondents indicate numerical information on a scale that is ascending or descending in nature. Frequency scales can be intervals, where the difference between points on the scale is equivalent (for instance, asking the number of days per week attending school: 1, 2, 3, 4 or 5) or they can be ordinal, where the difference between groups is not necessarily the same (for instance, asking respondents their age from 18 to 30, 31 to 50, 50 and above).

- **Ranking questions**, in which respondents are given a list of items and asked to compare items to one another and give items some sort of ranking based on an ascending or descending scale (eg rank which of these four restaurants you like better from 1 to 4, where 1 is most liked, and 4 is least liked).

- **Matrix style questions**, in which respondents have to fill in a value for each cell in a table (eg the number of female teachers and male teachers in a school for each year level). These questions are typically a burden for respondents and should be avoided where possible.

### 8.2.3 Layout and delivery

Presentation is a crucial element to questionnaire design. Materials that are laid out appropriately and efficiently are more likely to attract better quality responses in comparison to those that look disorganised and unclear. Questionnaires should be well laid out, should not be too cramped and should use a consistent sizing of font and structure should be used throughout. There should be sufficient space for respondents to enter their responses. Questions should not run over two pages unless this is unavoidable. If this has to happen, it is suggested to repeat the stem and response options again so respondents to not have to refer back to the previous page.

Typically some introductory text informing respondents about the nature of the questionnaire is included. The level of detail should be tailored to the respondents of the questionnaire. More detailed information would generally be given to a school head or head teacher than would be given to students. It is important that respondents are informed that the information they provide will not be identifiable back to the individual respondent, and that their privacy will be protected (assuming this is a feature of the questionnaire). It is usually customary to let respondents know that there is no right or wrong answer to the questions. It can be useful to give an indication as to the estimated amount of time that will be required to complete the questionnaire.

To avoid confusion within questionnaires with different question formats, each question should ideally have an instruction of how the respondents should respond. The instructions should be as concise as possible and only provide information necessary to respondents to answer the questions. An example, as shown in Exhibit 26 asks the respondents to mark an ‘X’ in one square only. Notes might also be used to give extra information or to impose conditions on the desired response.

Where data are to be manually entered (and not automatically scanned), it is suggested to use response option boxes that are numbered to ease the data entry process. It allows the person entering the data with the appropriate value to enter. So in Exhibit 26, if a student indicated that they were a boy, the person doing the data entry would enter a value of 2.

### Exhibit 26:
Example of a question using numbered boxes

I am a:

*Please mark an ‘X’ in one square only.*

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Girl</td>
<td>□1</td>
</tr>
<tr>
<td>Boy</td>
<td>□2</td>
</tr>
</tbody>
</table>
This chapter has discussed some of the important principles in designing tests and questionnaires. The ways in which items and tests are structured and laid out greatly affect the quality of the data that are collected. Therefore, it is important to consider aspects such as test length, the appearance of items within tests and the overall booklet design.

Questionnaires are often an important part of large-scale assessment programs. While tests provide information about learning outcomes, it is the data from questionnaires that put this information in context. Similarly to tests, questionnaires must be designed so that they collect relevant information and so that they are clear and attractive to respondents. For both test and questionnaire development it is important to set up thorough development processes, including robust quality control processes such as trialling the instruments.
# Checklist: Test design

If you are planning or implementing a large-scale assessment, have these points been addressed?

## Test design

1. Do the test(s) fulfil the assessment blueprint requirements (including test length and response formats)?
2. Have different linking designs been considered?
3. If using multiple test forms, have the forms been appropriately designed and linked according to data analysis requirements?

## Test layout and structure

4. Do the test booklets collect all relevant student demographic and test administration information?
5. Are appropriate instructions and practice items provided?
6. Have the items been ordered in the tests appropriately (considering item difficulties for example)?
7. Are all of the items in the test(s) laid out in a clear and uncluttered way?
8. Do all test items have labels printed next to them?
9. If using scored test items (closed constructed-response, short open-response or extended-response) have the scoring options been printed next to the items?
10. Has appropriate space been allowed for student responses and/or rater’s scores?
11. If multiple test forms have been used, has a document been created that lists all items including their unique label, the booklets they appear in, and the order in which they appear?

## Test proofreading and review

12. Have the test booklets been proofread before being printed?
13. Have the test booklets been checked after printing?
### Checklist: Questionnaire design

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Setting up the questionnaire design process</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have the broad areas of focus been selected for the questionnaire?</td>
</tr>
<tr>
<td>2. Have the interests of stakeholders been taken into account when selecting the content areas?</td>
</tr>
<tr>
<td>3. Has a decision been made about which questionnaires to develop, taking into account the positives and negatives of each approach?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Has the analysis and reporting of the questionnaire data been considered?</td>
</tr>
<tr>
<td>5. Have other potential sources for collecting contextual data been considered? (ie can the information be obtained from already existing sources?)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. If there is a student questionnaire, does it contain questions on background characteristics?</td>
</tr>
<tr>
<td>7. If there is a teacher questionnaire, will it be possible to match the teacher-level data to the student-level data?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reviewing and refining the questionnaires</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. Has the questionnaire been through a rigorous reviewing and editing process?</td>
</tr>
<tr>
<td>9. Has the questionnaire been piloted to see if the questions are working as intended?</td>
</tr>
<tr>
<td>10. Has qualitative feedback from the target population been gathered (where possible)?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Writing the questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Have all possible question types been considered when developing each question?</td>
</tr>
<tr>
<td>12. Are all questions clear and relatively short, and will they provide the data needed?</td>
</tr>
<tr>
<td>13. Are the questions appropriate for the target population, and could respondents answer them clearly and honestly?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layout and delivery considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>14. Is the questionnaire well laid out with sufficient space for responses and consistent fonts and structure throughout?</td>
</tr>
<tr>
<td>15. Will it be clear to respondents how they should respond to each question?</td>
</tr>
<tr>
<td>16. Does the questionnaire contain an introductory text on the purpose of the questionnaire and information related to confidentiality?</td>
</tr>
<tr>
<td>Glossary</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Blueprint</td>
</tr>
<tr>
<td>Cognitive processing skills</td>
</tr>
<tr>
<td>Common item linking</td>
</tr>
<tr>
<td>Common person linking</td>
</tr>
<tr>
<td>Construct</td>
</tr>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>Field trial</td>
</tr>
<tr>
<td>Greyscale</td>
</tr>
<tr>
<td>Horizontal linking</td>
</tr>
<tr>
<td>Items</td>
</tr>
<tr>
<td>Item cluster</td>
</tr>
<tr>
<td>Link unit</td>
</tr>
<tr>
<td>Link/anchor items</td>
</tr>
<tr>
<td>Raters</td>
</tr>
<tr>
<td>Response formats</td>
</tr>
<tr>
<td>Marking guides</td>
</tr>
</tbody>
</table>
### Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
<td>The part of the item that contains the question or task (e.g., in a multiple-choice item, the part that introduces the options).</td>
</tr>
<tr>
<td>Stimulus material</td>
<td>The prompt or context on which one or more items is based. For example, in a reading test, the stimulus is often a prose text made up of one or more paragraphs. In a mathematics test, the stimulus may be a diagram or a graph.</td>
</tr>
<tr>
<td>Target population</td>
<td>A particular group of people that the assessment is attempting to describe or for which it is attempting to measure outcomes. For example, an assessment may aim to measure reading ability of Class III students in government schools in a particular state. This group of people is referred to as the target population.</td>
</tr>
<tr>
<td>Test developers</td>
<td>Those responsible for producing test content, including contributing to the test design, writing items, helping to interpret trial data to select items for the main survey, and interpreting main survey data for reporting.</td>
</tr>
<tr>
<td>Test unit</td>
<td>A stimulus material and the item(s) that are based on the stimulus.</td>
</tr>
<tr>
<td>Vertical linking</td>
<td>The linking of items between tests administered to different class levels at the same time. Can be used to estimate growth between class levels (e.g., comparing Class III performance in 2015 to Class V performance in 2015).</td>
</tr>
</tbody>
</table>
References


Sample design
Sample design

Using a sample survey method in a large-scale assessment is an effective and efficient way of gathering information about the educational outcomes of groups of interest.

If sampling is to be used, it must follow well-established scientific procedures. This chapter outlines those procedures, starting with defining the groups of interest. Typically in educational assessments these groups will be made up of students who are within sections, that are in turn within schools, that are within districts. This chapter will outline the procedures for estimating the number of sampled students that will be needed. The chapter will also discuss how to organise the different levels of the sample – section, school, district, for example – to make the sample as statistically efficient and cost-effective as possible.
9.1 Purpose of sampling

Sampling is a technique used to select a subset (or sample) of elements from a population. This sample is studied in detail in order to estimate population attributes [1].

One of the main purposes of sampling, particularly in educational survey research, is to be able to make generalisations or inferences about the population from which the sample was drawn. This means that statements about the population can be based on the findings of the study conducted using the sample [1].

In order to make generalisations about the population of interest, it is imperative that the sample is selected using sound scientific procedures. This ensures that the sample is highly representative of the population and has the same characteristics as the population [1].

There are many advantages to studying a sample (a type of study known as a survey), as opposed to studying the entire population (otherwise known as a census) [1]. These advantages include:

- cost effectiveness – reduces costs associated with gathering and analysing the data, reduces requirements for trained personnel to conduct the fieldwork
- timeliness – faster data collection, data analysis and reporting
- human resources requirements – fewer human resources required, which means reduced costs involved in resources and training personnel
- quality control – possibly higher quality as it would most likely be easier to control testing conditions (more intense supervision of fieldwork and data preparation operations) to reduce bias.

However, as data are collected from a sample from the population of interest, the data are subject to ‘sampling error’ [2]. Sampling error occurs when the characteristics of the survey sample are somehow different from the characteristics of the population of interest. The larger the sampling error, the less representative the sample [2, 3]. There are two main types of error present when selecting a sample.

- error due to chance – because samples are a subset of the population of interest, there is always a chance that the true values resulting from a sample are not the same as the true values of the population [3]
- sampling bias – some systematic inclusion or exclusion process as a result of the selection method or due to respondents’ decision to participate or not participate in the study (also known as response bias) [2, 3].

Error due to chance can be minimised by increasing the sample size. However, sampling bias cannot be reduced or minimised by increasing the sample size. Instead, utilising a probabilistic sampling method (in which each population member has a chance of being selected and the probability of selection is known at the time of sampling [1, 4]) will help to reduce the bias.

If inadequate methods are used to select the sample resulting in a failure to minimise sampling error, the advantages of studying a sample to make generalisations and inferences about the target population are diminished [2, 3].
9.2 Defining the target population

It is important to clearly define the population that the survey is attempting to describe. A distinction is usually required between the desired target population and the defined target population. The desired target population is the population to which inferences from the survey outcomes will be made. The defined target population is the desired target population minus certain elements that are excluded due to practical difficulties [4]. As an example, the desired target population might be Class VIII students, but this desired target population becomes the defined target population when we remove elements of this population that may be difficult to cover [4]. Possible exclusions could include:

- Some Class VIII students are schooled at home, or are currently overseas.
- Some Class VIII students may be in institutions that are not generally classified as ‘educational’ – eg detention centres. Some institutions may not appear on the available list of schools used to as the sampling frame.
- Some Class VIII students no longer attend school or are ‘between schools’ at the time of the survey.
- Some Class VIII students may be unable to respond to the survey instruments – eg because insufficient facility with the testing medium of instruction, or physical or intellectual disability.
- Some Class VIII students are in schools that are difficult to access such that the costs for data collection are considered too high.
- Some Class VIII students in Indian schools will be exchange students from other countries temporarily being schooled in India. Is it intended that these students be included in the population?

For most surveys, the defined target population will differ from the desired target population. The portion of the desired target population who are not in the defined target population is referred to as the excluded population. Exclusions from the defined target population need to be clearly documented, and the degree of non-coverage needs to be estimated using available data sources – eg census and enrolment data [4].

Exclusions can be categorised as either whole-school exclusions (eg inaccessible17 schools; schools for children with intellectual disabilities) or within-school exclusions (eg students with physical or intellectual disabilities or limited language skills such that they are unable to participate in the assessment) [4].

It is preferable that the defined target population be as close as possible to the desired target population as the distinction between the two can tend to be overlooked in the discussion of the survey results [4]. As a general rule of thumb, the defined population should include at least 95% of the desired population. Exclusions, therefore, describe the difference between the desired and defined target populations and should be no more than 5% of the defined target population. The figure of 5% is best practice but whatever level is determined appropriate should be documented in the technical standards (see Chapter 4).

---

17The definition of inaccessibility depends on the context of the assessment project. For example, inaccessibility may be due to geographic, political, or security reasons. In addition, the standards for these reasons will vary depending on the clusters within the target population. For example, the standard for exclusion due to geographic remoteness may be different across states.
Exhibit 27 illustrates the relationships among the levels of the target populations and the possible reductions in coverage and exclusions. Ultimately, the defined target population is the population that the sample of participating students effectively represents, after all sources of exclusions have been taken into account [5].

**Exhibit 27:**
Coverage and exclusions

Exclusion criteria both at the whole-school and within-school level should be operationally defined to ensure accurate identification of the excluded population [1]. For example, very small schools were identified and excluded from the National Achievement Survey (NAS) Class X survey. This exclusion was operationally defined as ‘schools with fewer than 9 students taught in the language of the test in Class X’.[18] An example of an operationalised definition of within-school exclusion of students with insufficient language skills was defined as ‘if a student has moved to a school/location and has only then started in the test language, less than one year before the test, that student should be excluded’.

---

[18] This criterion for small schools is particular for NAS Class X. Other assessments may have slightly different cut-off values for the minimum number of students to be considered as ‘small school’ in the context of exclusion criteria.
9.3 Defining the sampling frame

A sampling frame is a list of all the sampling elements for a sample survey. A well-constructed sampling frame is one that provides complete coverage of the defined target population. For example, for an educational assessment survey of Class VIII students in a particular state, the sample frame will contain a list of all the schools in the state containing Class VIII students, along with an estimate of the number of Class VIII students in the school. Care should be taken during the construction of the sampling frame so that it excludes incorrect entries, duplicate entries or entries that refer to elements that are not part of the defined target population. The quality of the sampling frame has a direct effect on the survey results as it is used to ultimately select the subset (or sample) for the study; thus, frame construction should be treated with extreme care [1, 4].
9.4 Stratification

To improve the efficiency of the sample design and to ensure that all parts of the population are included and represented in the sample, schools can be grouped or ordered according to some stratification criteria. Stratification can be achieved explicitly and implicitly [4, 6]. Explicit stratification is a procedure whereby the sampling frame for a population is divided into separate sub-population frames, and then a separate sample is drawn from each sub-population. Explicitly stratifying the frame enables some control on the sample selection process by ensuring that a sample is spread over the distributions of the stratification variables that are used [4, 6].

As an example, if school type (government, government aided, private) is an explicit stratification variable, then separate school sampling frames would be constructed for each school type. The samples for each type would then be selected separately, possibly applying different sample designs [4, 6].

In practice, the most common reason for considering explicit stratification is to implement a disproportionate allocation of the school sample to the explicit strata. For example, the same number of schools could be sampled from each explicit stratum, regardless of the relative size of each stratum. The objective would be to produce equally reliable estimates for each school type. This differs from a proportional allocation where large strata would have more sampled schools than small strata. With a proportional allocation, the sample size is often too small in small strata to obtain estimates that are sufficiently reliable for comparison [4, 6].

In addition, implicit stratification can be applied to gain further control over the distributions of other variables. Implicit stratification refers to the sorting of the frame on other variables within each sub-population or explicit stratum. It is a very simple way of ensuring a strictly proportional sample allocation of elements/schools across all implicit strata [4, 6]. As an example, if school gender composition (female, male and co-educational) was used as an implicit stratification variable, the number of schools selected for each school gender composition would be about the same as what is observed in the population. That is, if the proportion of female schools was 20%, male schools was 65% and co-educational schools was 15%, sorting or listing the schools by gender composition prior to sampling will yield a drawn sample consisting of schools of gender composition similar to that of the population.

Stratification is generally used for the following reasons [4, 6]:

• to improve the efficiency of the sample design, thereby making survey estimates more reliable
• to apply different sample designs, such as disproportionate sample allocations, to specific groups of schools, such as those in different language groups
• to ensure that all parts of a population are included in the sample
• to ensure adequate representation of specific groups of the target population in the sample
• to obtain adequate representation of specific groups of the target population in the sample, if required.

Examples of stratification variables include:

• regions (states, provinces)
• urbanisation (rural areas, urban areas)
• socio-economic status (low, medium or high income)
• school types (government, government aided, private)
• school size.

Prior to sampling and preferably during the sample frame construction stage, thought should be given to the types of stratification variables required so that the data can be collected at the time of constructing the sampling frame [4, 6].
9.5 Multi-stage sampling

Multi-stage sampling, or cluster sampling, is a method that is commonly used in large-scale educational surveys. The population is divided into mutually exclusive units called clusters and a sample of clusters is selected to participate in the survey. The sampling of clusters can occur at multiple stages. That is, a number of schools can be selected first, then a number of sections within each sampled school and finally all or a subsample of students within the sampled sections. For very large populations, an area sampling stage – where geographical locations, such as districts, are selected first – can also be incorporated [7].

One of the advantages of using a clustered sampling design is that it is cost-effective [1, 7]. A larger group of students from the same school can be surveyed at the same time rather than possibly one or two students if a simple random sample design had been employed. This can result in saving on administrative costs [4]. Another advantage of using a clustered sampling design is that it allows for multilevel analyses of data, where the level of the school or the section within the school can be incorporated into the survey analysis [4].

Cluster samples tend to require a substantially larger sample size to achieve the same level of precision as a simple random sample (see Section 9.6). This is because students within clusters tend to exhibit more homogeneous (similar) characteristics than are observed in the population as a whole. This clustering effect can be attributed to all sorts of factors: for example, school environment, school location and school socio-economic intake [4, 7].

This clustering effect can be measured by examining the relative variance of outcomes within a cluster in relation to the variance in the population. The clustering effect is referred to as the intraclass correlation [4, 7].

The use of multi-stage sample designs that incorporate an area sampling stage where geographical locations are selected at the first stage can offer considerable administrative advantages. There is no need to visit all geographical areas to administer the survey. However, such an approach may lead to a loss of representation of sub-population groups if they are not evenly distributed across all geographical areas. Exhibit 28 provides a visual hypothetical illustration of this effect. Exhibit 28(a) shows a state divided into 10 districts. Exhibit 28(b) shows the clustering of 10 sampled schools selected within 5 sampled districts. Exhibit 28(c) shows the random spread of 10 selected schools within the state. As can be seen in this exhibit, when selecting geographical areas at the first stage, the resulting sample is less likely to be representative of the whole state. Such a sample design can be made more representative by increasing the sample size and using stratification.
Exhibit 28:
Visual example of the effects of clustering using a multi-stage design

<table>
<thead>
<tr>
<th>#</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| a. | State divided into 10 districts, colours represent student subgroups, or population subgroups |
| b. | State where 5 districts have been sampled and 2 schools have been selected within each sampled district |
| c. | State where 10 schools have been selected within the state |
9.6 Determination of sample size

The sample size is an important feature of any survey research which involves making inferences about a population from a sample. In practice, the sample size used in a survey is determined based on the cost and logistics of data collection, the need to have a sufficient level of precision and the analytical goals of the survey [1, 4].

Before sampling, it is important to determine a level of precision that is wanted around the main survey estimates [3]. As a rule of thumb, a precision level of +/- 5% around any statistic is a suitable goal. Generally to achieve this level of precision, a simple random sample of 400 elements is sufficient. However, in educational assessments, simple random samples are impractical. It is impossible to obtain a list of all students in the population and randomly sample 400. Even if this could be done, it is not practical to go to 400 locations to administer a survey just to a single student. Therefore, as noted above, a clustered sample is usually used. However, the use of clustering decreases the efficiency of the sample. This means that more students are required than would be the case under a simple random sampling scenario to achieve the same level of precision. The effect of using a clustered sample design is called the design effect.

9.6.1 Design effect

In order to calculate a sample size for a given survey, the type of sample design used and the design effect need to be taken into consideration. The design effect is an adjustment used in complex survey methods such as cluster sampling to assist in determining the study’s sample size. It is a measure of the extent to which a sampling design deviates from a simple random sampling design. It is often estimated in advance of sampling by reference to previous studies of a similar design [2, 4].

To calculate the design effect of a clustered sample, the following formula is used:

\[ \text{Deff} = 1 + (C-1) \times r \]

Where:
- \( \text{Deff} \) = Design effect
- \( C \) = the size of the cluster (number of students within the school who will be assessed in a subject)\(^{19}\)
- \( r \) = intraclass correlation\(^{20}\)

For example, a study has the following properties:

- Cluster size is 5, and using the estimated intraclass correlation of 0.6, therefore:

\[ \text{Deff} = 1 + (5-1) \times 0.6 = 3.4 \]

To calculate the equivalent to an effective sample size (ESS; that is, one that is equivalent to a simple random sample size) of 400, multiply 400 by the design effect to give the overall clustered sample size required (CSS).

\[ \text{CSS} = \text{ESS} \times \text{Deff} \]

\[ \text{CSS} = 400 \times 3.4 = 1360 \]

So given the cluster size and the intraclass correlation values above, 1360 students, in clusters of 5, would be required to obtain the equivalent of 400 students using simple random sampling.

\(^{19}\) The cluster size is determined by the number of students who will be assessed in each subject.

\(^{20}\) The intraclass correlation is a descriptive statistic that can be used when quantitative measurements are made on units that are organized into groups. It describes how strongly units in the same group resemble each other.
The effects of clustering need to be estimated and taken into account when determining the appropriate sample size. Clustering can be cost-effective (e.g., savings on logistic costs) but increased clustering has the effect of increasing the required sample size [6]. A balance needs to be carefully considered when designing the sample and deciding on the extent of clustering.

It is important to note that the accuracy of the sample depends on the sample size or number of observations and not on the size of the sample relative to the population from which the sample is drawn [1, 2, 8]. For example, a sample of 1000 from a population of 1 million (with a sampling fraction\(^2\) of 0.001 or 0.1%) is much better than a sample of 100 from a population of 10,000 (although the sampling fraction for this sample is 0.01 or 1%).

\(^2\)The sampling fraction is calculated using the formula \(n/N\) where \(n\) = sample size and \(N\) = population size.
9.7 Field trial sample design

There are usually two data major collection stages involved in large-scale surveys, the field trial and the main survey. The primary aims of the field trial, otherwise known as a ‘pilot study’, are to test the survey instruments and to test operational procedures. Analysis of the field trial data is undertaken to check that the survey items are performing correctly in measuring the outcomes of interest. The experience with the field trial operations is used to refine instruments and procedures for the main survey.

Because the field trial results are not publicly reported, the sampling approach for the field trial does not need to be as rigorous as for the main survey. The trial might be restricted to a limited number of states or districts in order to contain costs and to minimise the burden on smaller jurisdictions. However, it is desirable that the selection of schools for the field trial be approximately representative of the range of different school types that occur within the population of interest. For example, the survey should cover different sectors and geographical locations.

The size of the field trial sample is based on the amount of response data required to be able to adequately test the psychometric properties of the survey items. Usually 200 responses per item are considered sufficient for this purpose. Often the field trial will include many more items than are expected to be carried through to the main survey, possibly spread over several forms. A sampled student completing one of these forms will only be providing responses to a subset of all of the available items. A larger number of students may therefore be necessary to achieve the desired number of responses per item. The items that perform best according to the data analysis and the set of items that best cover the survey framework are selected for the main survey.
9.8 Replacement protocols

It is not always possible to obtain the participation of all sampled schools. In order to avoid the resulting sample size losses and minimise school and student non-response, a mechanism to identify a priori replacement schools for non-participating sampled schools is needed [4, 5]. This means that replacement schools should be identified even before knowing (a priori or before the fact) which sampled schools will not participate.

An important reason for identifying replacement schools a priori is to avoid the haphazard use of alternate schools as replacements, which may amplify response biases. Although an a priori approach does not necessarily avoid non-response bias, it tends to minimise the potential for bias. Furthermore, it is conceptually more desirable than over-sampling to accommodate a low participation rate [4, 5].

Replacement schools are identified at the time of sampling by assigning the two schools neighbouring the sampled school in the frame as substitutes to be used in instances where an original sampled school refuses to participate. Replacement schools are required to be in the same implicit stratum (ie have similar demographic characteristics) as the sampled school [4, 5]. For example, if a sampled school is a government school, located in a rural area and has Hindi as the medium of instruction, the predetermined replacement schools should also be a government school located in a rural area and have Hindi as the medium of instruction.
Within-school sampling methods refer to the selection of students within schools. It can involve direct sampling of students in the targeted class at the sampled school or it can involve the sampling of intact section(s). If sampling intact section(s), either all students within the sampled section(s) can be tested, or an additional sampling level can be applied in which a subsample of students is selected within the sampled section(s).

Within-school sampling methods are usually conducted on the day of testing. If direct student sampling is used, a list of all students in the target class is obtained from the participating school and the students are randomly selected from the list. If the sample design requires intact sections, a list of sections is obtained from the participating school, and a section or sections are then selected. An additional sampling step may be carried out to select a subsample of the students within the sampled section.

Regardless of the number of levels of within-school sampling, it is critical that the selection method is well-documented and the selection process is carried out the same way across all sampled schools. Any deviation from the prescribed sample selection process could compromise the sample by introducing potential bias.

In addition to ensuring consistency across the within-school sampling activity, the following information relating to within-school sampling should be retained for each and every sampled school to ensure that all data required for the weighting task can be carried out:

- the list of sections and number of students within those sections
- the actual calculations used to sample the section(s)
- the list of students within the sampled section(s)
- the actual calculations used to sample the students (if applicable).

In the absence of documentation of the within-school sampling process, no quality assurance checks will be possible and the benefits listed below will be lost or compromised:

- consistency across the school’s sampling activities
- providing an essential reference for sample weighting (see Section 9.9), analysis and database construction
- providing information for technical reporting
- providing a reference for future cycles of a similar study.

Failure to document and retain the within-school sampling process of sections and/or students could have serious implications for the data analysis and reporting phase. For example, if no clear documentation was retained of the within-school sampling procedures, this would make quality control assurance difficult. Inconsistent collection and retention of section and student lists would mean that important data required to perform the weighting task could be compromised. Weighting may not be performed at an optimal level due to omission or failure to collect important data.
Timing of the within-school sampling process should also be taken into consideration. While the within-school sampling procedures can be carried out on the day of testing under certain circumstances, these activities would benefit from being conducted in advance.

For example, the sampling frame may contain parent schools: schools that act as a ‘parent’ to a cluster of other schools (child schools), with all examination enrolments going through the parent school. For sampled schools that are parent schools, it is important that parent and child schools are identified well in advance, so that section and student subsampling (where applicable) can be carried out in a timely manner. This is important to allow for cross checking of enrolment lists to ensure all sections from all child schools have been obtained. If sections from the child schools are left out of the within-school sampling process, this will compromise the sample by introducing potential bias. It will also compromise the sample weighting, particularly if it is not well-documented so that post-hoc weighting treatments can be applied. Additionally, as child schools can be close in proximity or some distance away from the parent school, within-school sampling should be carried out well in advance, so that if a section from a child school is sampled, testing can be carried out at that child school.
9.10 Response rates

Response rate is a basic measure used to evaluate the quality of the collected data. It is calculated as the number of elements that participated divided by the number of elements that were sampled expressed as a percentage [2].

\[
\text{Response rate} = \frac{\text{number of sampled elements that participated in survey}}{\text{number of elements that were sampled}} \times 100
\]

For example, if 1000 students were sampled and 800 students participated in the survey, then the response rate is calculated as:

\[
\text{Response rate} = \frac{800}{1000} \times 100 = 80\%
\]

The response rate can be used to gauge the potential for non-response bias. Non-response bias is the error resulting from distinct differences between students who participated in the survey versus the students who did not participate.

Non-response bias can have a negative effect on the generalisability of survey results. The lower the response rate, the higher the level of error. This creates a risk of the sample becoming less representative of the target population and may lead to misleading results, which in turn may lead to incorrect inferences about the target population.

For example, to see the effect that non-response bias might have on survey results, four hypothetical surveys are presented in Exhibit 29. The hypothetical survey was conducted using the same 10 students to find out the percentage that received A's in a recent test.
This example illustrates how non-response bias can affect how well data represents the population being surveyed. Relying on Survey #1 data for decision-making is highly problematic. Survey #2 is much closer to the actual, but is still not a very good estimate. Survey #3, though, with a higher response rate, gives a fairly close estimate of the true value - only 3 percentage points off.

The higher the response rate of a survey, the lower the risk of non-response bias. In large-scale educational surveys it is best practice to aim for a school response rate of at least 85% and a student response rate of 85%.

Every effort should be made to ensure a high number of sampled schools and students participate in the survey. Efforts such as promoting the importance of the school’s participation and the utilisation of replacement schools can raise the school response rates. Conducting follow-up testing administration sessions when participation rates of test administration sessions are low will assist in raising the student-level response rates.
9.11 Sampling weights

Sampling weights are used to correct for imperfections associated with the sample that might lead to bias and other departures between the sample and the population of interest. The purposes of weighting are [9]:

- to account for the sample design – compensates for any stratification or disproportional probabilities of selection of subgroups
- to adjust for non-response and non-coverage (due to errors in the sampling frame) of the population
- to adjust the weighted sample distribution for key variables of interest (eg age, sex)
- to make it conform to a known population distribution.

Once the imperfections in the sample are compensated for, weights can then be used in the estimation of population characteristics of interest and also in the estimation of the sampling errors of the survey estimates generated [9].
9.12 Summary

Using a sample survey method for a large-scale assessment saves money and time. A carefully
designed and executed sample will yield valid and reliable information, provided response
rates are sufficient. The population of interest must be clearly defined. A suitable scientific
sampling method must be used, an appropriate sample size determined, and precautions
taken to reduce sampling error, such as the use of stratification. Replacement protocols and
within-school sampling methods require consideration at the time of sampling.

It is essential that every step of the sampling design and implementation process is well-
documented. Failure to fully document within-school sampling, for example, can lead to
erroneous assumptions about lack of potential non-response bias in the sample.
Checklist: Sample design

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Purpose of sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have cost and time constraints been factored into the sample design?</td>
</tr>
<tr>
<td>2. Has thought been given to minimising sampling error (e.g., use of a scientifically sound selection procedure)?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Defining the target population</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Has the desired target population of interest been identified and clearly described?</td>
</tr>
<tr>
<td>4. Has the defined target population been clearly described?</td>
</tr>
<tr>
<td>5. Have all elements that DO NOT form part of the desired target population - that is, all school- and student-level exclusions - been defined?</td>
</tr>
<tr>
<td>6. Have all exclusion criteria for whole schools been operationally defined?</td>
</tr>
<tr>
<td>7. Have all exclusion criteria at the within-school level been operationally defined?</td>
</tr>
<tr>
<td>8. Has the degree of non-coverage been estimated using available data sources and documented?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Defining the sampling frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Is there a sampling frame available? If not, has a sampling frame that contains ALL elements that make up the defined target population been constructed?</td>
</tr>
<tr>
<td>10. Have all elements that DO NOT form part of the desired target population been identified, removed from the sampling frame and documented?</td>
</tr>
<tr>
<td>11. Does the sampling frame include all elements of the defined target population?</td>
</tr>
<tr>
<td>12. Does the sampling frame include key elements of interest (stratification variables, school-level information, and so on)?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stratification</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Have all explicit stratification variables been identified?</td>
</tr>
<tr>
<td>14. Have all implicit stratification variables been identified?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multi-stage sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>15. Have multi-stage sampling methods been considered?</td>
</tr>
<tr>
<td>16. Have the clustering effects of multi-stage sampling been considered?</td>
</tr>
</tbody>
</table>
### Determination of sample size

17. Has an acceptable level of statistical precision been chosen?  
18. Has the sample design effect been calculated and the sample size determined?

### Field trial

19. Does the field trial sample approximately represent the range of different schools that the population of interest attend?  
20. Is the sample size of the field trial calculated based on the response data required to be able to adequately test the psychometric properties of the items?

### Replacement protocols

21. Have protocols for replacing schools at the time of sampling been developed?  
22. Have all schools been assigned implicit stratification variables?

### Within-school sampling

23. Has a method for selecting students been decided upon?  
24. Has the timing (on the day of testing or some time in advance of the testing day) of within-school sampling been determined?  
25. Has a method been devised to fully document all within-school sampling information?
## Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster sampling</td>
<td>A sampling technique used when ‘natural’ but relatively homogeneous (similar) groupings are evident in a population of interest.</td>
</tr>
<tr>
<td>Design effect</td>
<td>An adjustment used in complex survey methods to assist in determining the study’s sample size.</td>
</tr>
<tr>
<td>Desired target population</td>
<td>The population to which inferences from the survey outcomes will be made.</td>
</tr>
<tr>
<td>Defined target population</td>
<td>The desired target population minus certain elements that are excluded due to practical difficulties.</td>
</tr>
<tr>
<td>Effective sample size</td>
<td>The sample size that would be used under a simple random sample design.</td>
</tr>
<tr>
<td>Excluded population</td>
<td>The population that does not form part of the defined population.</td>
</tr>
<tr>
<td>Explicit stratification</td>
<td>A procedure whereby the sampling frame for a population is divided into separate sub-population frames, and then a separate sample is drawn from each sub-population.</td>
</tr>
<tr>
<td>Implicit stratification</td>
<td>The sorting or listing of variables of interest to ensure a strictly proportional sample allocation of elements during sampling.</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>A descriptive statistic that can be used when quantitative measurements are made on units that are organised into groups. It describes how strongly units in the same group resemble each other.</td>
</tr>
<tr>
<td>Field trial</td>
<td>Administration of items under test conditions, used to test the items’ validity and the administration procedures. Occurs before the main survey with a sample of at least 200 students per item who are similar to the target population.</td>
</tr>
<tr>
<td>Sample frame</td>
<td>A list of all the sampling elements for a sample survey. For example, for an educational assessment survey of Class VIII students in a particular state, the sample frame will contain a list of all the schools in the state containing Class VIII students, along with an estimate of the number of Class VIII students in the school.</td>
</tr>
<tr>
<td>Variance</td>
<td>A numerical measure of how the data values are dispersed around the mean.</td>
</tr>
</tbody>
</table>
References


Materials production and field operations
Materials production and field operations

This chapter outlines the planning and preparation of resources required to successfully administer the tests during the field trial and main survey stages.

It is divided into three sections. The first (Section 10.1) concerns preparatory activities leading to the testing day; the second (Section 10.2) is centred around the activities on the day of testing; and the third (Section 10.3) is focused on post-testing activities.

There are two major data collection stages involved in large-scale assessment surveys – the field trial (otherwise known as the pilot study) and the main survey. The primary aims of the field trial are to evaluate the assessment instruments and the operational procedures. As a result of the field trial, any identified issues relating to the assessment tools or operational procedures can then be improved and implemented in the main survey. Therefore, the materials production and field operations activities that are carried out in the field trial are duplicated in the main survey with some adjustments where required.
10.1 Preparatory stage

10.1.1 Operation plan

The operation plan should contain a comprehensive list of activities, resources and the timelines to which they need to be completed. One of the benefits of having an operation plan is that it provides a detailed overview of the activities, resources and the timeline to which these need to be addressed and completed. This helps to ensure that the data are collected and are ready for analysis in a timely manner.

Another benefit is that the operation plan can be adjusted to accommodate any unforeseen delays. As the important activities and resources have been listed in the operation plan, this minimises the risk of accidently omitting activities if adjustments to timelines need to be made.

Exhibit 30 is an example of an operation plan for the different stages of an assessment project through to data collection stage. This operation plan lists when the activities and resources are to be delivered and to what stage of the project the activities and resources relate: the field trial (FT) or the main survey (MS).

**Exhibit 30:**
Example of an operation plan of an assessment project through to data collection stage

<table>
<thead>
<tr>
<th>Timeline</th>
<th>Activities and resources</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>Project manager manual draft</td>
<td>Project initiation</td>
</tr>
<tr>
<td>2013</td>
<td>FT sampling</td>
<td>Field trial (FT)</td>
</tr>
<tr>
<td>Feb</td>
<td>FT booklets, questionnaires and manuals translation/adaptation</td>
<td>Main survey (MS)</td>
</tr>
<tr>
<td>Mar</td>
<td>Field operations workshop</td>
<td>Analysis &amp; reporting phase</td>
</tr>
<tr>
<td>Apr</td>
<td>FT material preproduction and printing</td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>FT convenience sample constructed</td>
<td></td>
</tr>
<tr>
<td>Jun</td>
<td>FT schools contacted, information on classes and students obtained</td>
<td></td>
</tr>
<tr>
<td>Jul</td>
<td>FT operations manuals translated</td>
<td></td>
</tr>
<tr>
<td>Aug</td>
<td>FT field investigators sourced (identified) and trained</td>
<td></td>
</tr>
<tr>
<td>Sep</td>
<td>FT test material dispatch to relevant authority</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Finalise MS sampling plan</td>
<td></td>
</tr>
<tr>
<td>Jan</td>
<td>FT data submission</td>
<td></td>
</tr>
<tr>
<td>Feb</td>
<td>FT data cleaning</td>
<td></td>
</tr>
<tr>
<td>Mar</td>
<td>MS item translations reviewed and adapted</td>
<td></td>
</tr>
<tr>
<td>Apr</td>
<td>MS booklets, questionnaires and manuals adapted</td>
<td></td>
</tr>
<tr>
<td>Sep</td>
<td>MS sampling complete</td>
<td></td>
</tr>
<tr>
<td>Oct</td>
<td>MS schools contacted, information on classes and students obtained</td>
<td></td>
</tr>
<tr>
<td>Nov</td>
<td>MS operational manuals updated</td>
<td></td>
</tr>
<tr>
<td>Dec</td>
<td>MS material preproduction and printing</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>MS field investigators sourced and trained</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>MS data collection</td>
<td></td>
</tr>
<tr>
<td>Jan</td>
<td>MS test material dispatch to relevant authority</td>
<td></td>
</tr>
<tr>
<td>Feb</td>
<td>MS scanning complete</td>
<td></td>
</tr>
<tr>
<td>Mar</td>
<td>MS coding</td>
<td></td>
</tr>
<tr>
<td>Apr</td>
<td>MS data submission</td>
<td></td>
</tr>
<tr>
<td>Aug</td>
<td>MS data cleaning</td>
<td></td>
</tr>
<tr>
<td>Nov</td>
<td>MS data analysis</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>MS data analysis</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>MS reports</td>
<td></td>
</tr>
</tbody>
</table>
An operation plan may encompass all activities associated with the project, or alternatively, there can be several operation plans that are specific to key personnel (see Chapter 3 for the core project team members and their roles). For example, operation plans can be developed for project managers, school coordinators, field investigators and quality assurance monitors. Exhibit 31 is an example of an operation plan for the field investigator.

**Exhibit 31:**
Example of an operation plan for a field investigator

<table>
<thead>
<tr>
<th>Date / timeframe</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Before the assessment</td>
<td></td>
</tr>
<tr>
<td>Before training</td>
<td>1. Receive notice of assessment dates and times</td>
</tr>
<tr>
<td>&lt;Training date&gt;</td>
<td>2. Attend training session for field investigators</td>
</tr>
<tr>
<td>After the training</td>
<td>3. Familiarise yourself with explanatory notes on student questionnaire items</td>
</tr>
<tr>
<td>Two weeks prior to the assessment</td>
<td>4. Receive, check and secure package of assessment materials</td>
</tr>
<tr>
<td>One week prior to the assessment</td>
<td>5. Confirm assessment plans with school coordinator</td>
</tr>
<tr>
<td>II. Assessment day</td>
<td>6. Prepare all materials necessary for the assessment</td>
</tr>
<tr>
<td>Assessment sessions</td>
<td></td>
</tr>
<tr>
<td>Immediately after the assessment</td>
<td>7. Give school questionnaire to school coordinator</td>
</tr>
<tr>
<td>Assessment sessions</td>
<td>8. Set up room and materials</td>
</tr>
<tr>
<td>III. After the assessment</td>
<td>9. Consider important aspects of test administration</td>
</tr>
<tr>
<td>Soon after the assessment</td>
<td>10. List all students present in the class then conduct within-school sampling</td>
</tr>
<tr>
<td>As soon as all materials are ready for despatch</td>
<td>11. Conduct sessions and collect the test materials; record participation in student attendance sheet</td>
</tr>
<tr>
<td>As soon as all materials are ready for despatch</td>
<td>12. Append any important notes about the test session to field notes</td>
</tr>
<tr>
<td>As soon as all materials are ready for despatch</td>
<td>13. Collect completed questionnaires</td>
</tr>
<tr>
<td>As soon as all materials are ready for despatch</td>
<td>14. Conduct quality assurance, including manual scrutiny</td>
</tr>
<tr>
<td>As soon as all materials are ready for despatch</td>
<td>15. Determine if a follow-up session needs to be held</td>
</tr>
<tr>
<td>As soon as all materials are ready for despatch</td>
<td>16. Organise completed assessment materials for despatch (as per guidelines)</td>
</tr>
</tbody>
</table>

22 For NAS Class X, test materials were sent directly to the participating school. Best practice would be to assign responsibility of test materials to the field investigator to ensure the security of the test items for future use.

23 Within-school sampling is done on the assessment day for NAS Class X and State Level Achievement Surveys (SLAS). Some assessment projects may conduct the within-school sampling beforehand using specialised data management software or it can be done manually.
The production of materials involves several rounds and thought should be given to the time and resources required in order to produce all print materials in a timely manner. There are two main types of materials that need to be produced: assessment materials and administration materials.

Assessment materials include the test booklets, questionnaires (e.g., student, school, teacher and/or parent) as well as associated supporting documents such as answer or optical character recognition/optical mark recognition (OCR/OMR) sheets.

Administration materials include manuals relating to the administration of the tests (otherwise known as field guidelines or field operations manuals), as well as important supporting documents such as student attendance sheets (sometimes referred to as student tracking forms) and field notes, which should include information about within-school sampling.

Exhibit 32 lists some key administration documents that may be produced to assist with the preparation of resources to support the field operations tasks, listed by project stage: field trial (FT) or main survey (MS).

### Exhibit 32:
List of key administration documents to be produced in different stages of the survey

<table>
<thead>
<tr>
<th>STAGE</th>
<th>KEY DOCUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field trial</td>
<td>School coordinator’s manual FT</td>
</tr>
<tr>
<td></td>
<td>Field investigator’s manual FT</td>
</tr>
<tr>
<td></td>
<td>Student attendance sheet FT</td>
</tr>
<tr>
<td></td>
<td>Field notes FT</td>
</tr>
<tr>
<td>Main survey</td>
<td>School coordinator’s manual MS</td>
</tr>
<tr>
<td></td>
<td>Field investigator’s manual MS</td>
</tr>
<tr>
<td></td>
<td>Student attendance sheet MS</td>
</tr>
<tr>
<td></td>
<td>Field notes MS</td>
</tr>
<tr>
<td></td>
<td>Quality monitor’s manual MS</td>
</tr>
</tbody>
</table>

#### 10.1.2.1 Developing the administration manuals

The administration manuals are documents that provide clear instructions as to the role and core tasks key personnel such as the school coordinator, field investigators and quality monitors undertake to ensure that the administration of the tests are carried out smoothly and in a standardised way across all participating schools.

The school coordinator and field investigator manuals need to be prepared for the field trial and amended to include any procedural changes identified at the field trial stage for the main survey. The quality monitor’s manual need only be produced for the main survey stage.

For each administration manual, the structure should contain the following information:

- a brief overview of the purpose of the survey and the way the data will be used
- details of the role and core responsibilities of personnel (specific to the role each manual targets; i.e., the school coordinator, field investigators and quality monitors)
- a description of the resources and preparation that are needed for tasks
- the contact details of relevant authorities.
Examples of some of the core tasks to include in each administration manual are outlined below.

Core tasks in the school coordinator’s manual include:

• acting as the liaison between the schools, the institutional coordinator and/or pedagogical coordinator and the field investigator
• compiling section and/or student lists
• notifying teachers, students and parents of the assessment
• organising and confirming assessment plans with school staff and with field investigator
• assisting the field investigator in organising the assessment activities (setting up rooms and materials)
• checking that the field investigator has correctly completed the student attendance sheet at the completion of test administration.

Core tasks to be described in the field investigator’s manual include:

• attending training session for field investigators
• carrying out the within-school sampling plan (sampling of sections and or students)
• assigning test booklets to sampled students
• administering the testing session (include rules for assisting students and scripts to be read out)
• recording the timing of test sessions
• completing a student attendance sheet for each test session
• recording student exclusion or participation codes using a pre-defined list
• determining whether a follow-up session is needed
• organising the assessment materials for despatch
• returning the assessment materials to relevant authority.

The quality monitor’s manual should include the testing procedures and defined standards, and a checklist to record or assess whether the standards have been met in terms of:

• steps taken to carry out the within-school sampling
• documentation for all within-school sampling is attached or copied onto the field notes
• the field notes has been completed with no missing data
• recording of attendance
• distribution of test materials to students
• time taken for the test
• how closely the field investigator followed the instructions defined in the field investigator manual
• collection of test material
• students’ behaviour
• any unpredictable event or additional observatory notes.

Additionally, core tasks to be described in the quality monitor’s manual include:

• attending training session for quality monitors and field investigators
• carrying out checks on the calculations used for within-school sampling and documenting the extent to which testing procedures in schools meet the defined standards
• performing within-school sampling so that it can be checked thoroughly
• observing the test (it should be made clear that the quality monitor is an observer of the testing and is there to record this observation and should not hinder or disrupt the testing in any way)
• returning quality monitor checklists to relevant authority.
For each of these core tasks, clear succinct step-by-step instructions as to how they are to be carried out should be included. These step-by-step instructions could include fine details such as applying coding schemes or the inclusion of a standardised spoken script to ensure uniform delivery of instructions across all participating schools.

In addition to the key administration manuals, supporting documents such as the field notes and student attendance sheet also need to be created. These supporting documents are essential to the data entry, analysis and quality assurance stages.

10.1.2.2 Developing the field notes

The field notes document is used to aid the collection of important school-level information and should include within-school sampling data. The completion of the field notes should be part of the field investigator's role.

Failure to include the collection of these data could mean the loss of information that is essential to accurately carry out the sample weighting. It also means that the within-school sampling process cannot be verified during the quality monitoring stage.

The field notes document should enable the collection of the following data for each participating school:

- school identification codes (one of particular importance is the identification code that links the school back to the sampling frame)
- key characteristics of the school, for example, school gender (male, female, co-educational), school type, urbanicity (urban/rural), management of school, affiliating board, medium of instruction
- school sampling status – that is, original sampled school, first or second replacement and reason for replacement if it is a replacement school
- school enrolment details such as:
  - the list of all sections of target class and the number of students enrolled in each target class section
  - the number of sections sampled
  - the calculations used to sample the section(s); that is, the random number and sampling interval (if more than one section was sampled)
  - the list of students for the sampled section(s)
  - the number of students selected, the random number and the sampling interval if a subsample of students was selected within the sampled section
  - notes relating to the test administration session.

It should be noted that if the list of sections and/or students is long and the school roll call or student registry was used on the day of testing, it is acceptable to include a photocopy of the original lists in place of retyping or rewriting this information on the field notes. Care must be taken to ensure that all required information is included should this be the case.
10.1.2.3 Developing the student attendance sheet

The student attendance sheet (otherwise known as the student tracking form) records student-level information and should include the following:

- school identification variables (this ensures that the student-level information can be linked to the school-level information)
- list of sampled students
- students’ names
- students’ assigned identification numbers
- the test booklet each student was assigned
- the students’ participation status (that is, whether they participated or were excluded based on pre-defined exclusion criteria – this should be completed by the field investigator on the testing day).

A separate student attendance sheet should be completed by the field investigator for each test session.

10.1.2.4 Quality assurance of materials

It is important that quality control methods are implemented to ensure assessment and administration materials are of high quality and are created in a timely manner. Quality assurance of materials production can be applied to ensure that all printed materials are produced to meet specific requirements or specifications. All print materials should undergo an editing stage prior to their mass production. The editing stage should include proof reading and checks to ensure the following [1]:

- required content is included
- spelling and typographical errors are removed
- font size and spacing of lines of text are of readable level
- diagrams are simple and clear, and where possible appear on the same page as the relevant text
- numbering of pages and margin alignment are correct.

Once all print materials have undergone the editing stage, a complete version of each document should be printed for a final optical check. The final optical check should confirm that each printed document will be of the same standard with respect to:

- the print quality (the same level of darkness in ink)
- the paper quality (the same type of thickness, weight and colour of paper).

Once the print materials have been finalised and a printing company has been sourced, a printing plan should be devised to include:

- print quantities:
  - the number of copies to be printed for each test booklet, supporting documents (eg answer sheets), questionnaires (student, school, teacher and parent if applicable), administration manuals (district coordinator, field investigator, quality monitor) and record-keeping materials (field notes and student attendance sheets).
- packing and collation instructions:
  - clear instructions outlining how the test booklets are to be assembled
  - clear instructions outlining how many of each print material should be packed for a state or school.
- distribution or delivery of materials:
  - timeline outlining where the materials need to be delivered and the quantities that need to be delivered.
- quality assurance plan that includes the proportion of print materials that should be checked to ensure they meet:
  - the printing quality standards
  - packing standards
  - distribution timelines.
10.1.3 Field operations training

Training facilitates the communication of roles and responsibilities key personnel have in the test administration process.

Before the training of personnel can begin, recruitment should take place. Recruiting of key school-level personnel needs to be done before both the field trial and main survey stages.

Key school-level administrative personnel include school coordinators, field investigators and quality monitors. In India, it is common to have district coordinators that are responsible for the participating schools within the district. It is international best practice to have school coordinators in addition to (or even instead of) district coordinators. School coordinators should be appointed for each participating school both at the field trial and main survey stages. Ideally, each school coordinator should be a staff member from the participating school [2].

If a district coordinator is appointed instead of a school coordinator, care should be taken to ensure that all the responsibilities listed for a school coordinator at the school level (e.g., compiling section and student lists, assisting to organise the assessment by setting up the room and materials, being present at the completion of the testing session to immediately conduct quality assurance tasks) are reassigned to another school-level key staff member.

Field investigators will need to be recruited to administer the assessment at participating schools for both the field trial and main survey. The number of field investigators will depend on the number of sampled schools and how many test administrations they will be responsible for.

Quality monitors should be appointed for the main survey stage only. The number of quality monitors will depend on the number of schools to be visited. The distribution of relevant administration manuals to school coordinators, field investigators and quality monitors should ideally occur before training begins at each stage. However, if this not be possible, this should occur no later than four weeks before the assessment. This will allow enough time for the school coordinators, field investigators and quality monitors to review the manual and complete any preparatory tasks required for the assessment [2, 3].

The training for all key school-level administration personnel should include the purpose of the test and their roles and responsibilities in its administration. School coordinators, field investigators and quality monitors should understand why standardised test administration procedures are vital, and thus, particular emphasis should be placed on uniform testing conditions during training [2].

Training of both school coordinators and field investigators needs to take place for both field trial and main survey stages. Training for quality monitors only needs to occur before the main survey stage.

Separate training sessions or one training session can be held for all school coordinators,
field investigators and quality monitors. The contents of the training session(s) could include the following [2]:

- an introduction - description of the purpose of the assessment and how the results are used
- description of the purpose and key responsibilities of each role, emphasising the importance of uniform testing conditions
- review of activities before, during and after testing as detailed in each administration manual
- review of the student attendance sheet and procedures for their completion, including some practice examples
  - review of specific guidelines to which key personnel should adhere:
    - field investigators should be aware of the within-school sampling procedure and adherence to scripts, recording of session times and rules governing the level of assistance they can render to students
    - quality monitors should know how to check the calculations for the within-school sampling, understand their role as observers and should not hinder or disrupt the testing in any way
- review of the explanatory notes to be used as a reference during the administration of the student questionnaire (also sometimes referred to as the pupil questionnaire)
- review and discussion of issues which commonly arise, such as managing students, admitting students who are late, recording students who leave the session temporarily, managing disruptive students
- discussion about the security of materials at all times of test administration and the necessity for the swift return of all assessment materials
- review and discussion of any specific issues, such as protocols for entering schools or communicating with school staff [2].

10.1.4 Contacting schools

Once the selection of schools has been finalised, all sampled schools need to be contacted and invited to participate in the assessment program. Permission to invite schools may require the permission of higher governing bodies beforehand.

With the approval of the relevant governing bodies, a letter should be sent to the school heads requesting their schools’ participation. The letter should be signed by the project coordinator/manager. The letter should [3]:

- include a description of the purpose of the assessment program, including a short document describing the project and its significance
- invite the cooperation of the school and emphasise the importance of the school’s participation for achieving a representative sample
- outline what participation will involve for the school (eg the class and number of sections and students involved, the length of the testing sessions, the approximate time required for the completion of the school and/or teacher questionnaire)
- give the proposed dates during which the testing will take place
- indicate any benefits that the schools might receive from participating in the study (eg feedback on school data)
- guarantee the anonymity of the individual students, teachers and schools in the publications of results
- provide school heads with contact details of the relevant authority if they need further information before making a decision
- ask for the name of an individual who could serve as the district coordinator and be responsible for the administrative arrangements for the project in the district
- ask for the name of the person to whom future correspondence should be addressed.

To minimise the risk of non-response bias, it is important to make every effort to ensure that
all schools in the list of sampled schools cooperate. Therefore, the letter should be sent to schools as soon as possible to maximise time to contact replacement schools should the need arise.

If a school is unable to participate, it must be replaced with the replacement school from the first replacement sample that has been pre-assigned to this school. If the first replacement school is also unable to participate, it should be replaced by the corresponding pre-assigned second replacement. If the second replacement school is still unable to participate, the sampled school is declared as non-participating. It must be noted that under no circumstances should a sampled school be replaced with a replacement school other than the two that were pre-assigned during the school sample selection process.

Once schools have agreed to participate, it is important to work with the schools to ensure that staff and students are engaged and cooperative. Continue communicating with schools to ensure they are informed about all aspects of the test administration.

10.1.5 Preparing Resources

The construction and distribution of assessment and administration materials are of key importance. The printing and distribution of assessment and administration materials must be done in a timely manner to ensure the receipt of materials occurs well in advance of the planned administration day to ensure the test is administered within the scheduled timeframe.
10.2 Implementation stage

10.2.1 Test administration

Test administration refers to the activities carried out at the school level and includes the receipt of materials, sampling of sections and/or students, administration of the tests, quality assurance monitoring, and the returning/collection and storing of the completed materials. Prior to arriving at the school on testing day, the field investigator will need to prepare and take the following materials to the school [2]:

- the field investigator’s manual
- student attendance sheet
- booklets containing the assessment material and student questionnaire
- school questionnaires and teacher questionnaires (the school coordinator should arrange for these to be completed by relevant people at the schools)
- pencils, erasers and a pencil sharpener
- a watch, a clock or other timing device.

Note that while test materials were sent directly to participating schools for the National Achievement Survey (NAS) Class X, it is best practice to assign the receipt and responsibility of test materials to the field investigators. This ensures the security of the testing materials for future use. The school coordinator and other school staff members should NOT be allowed to look through test booklets or copy or photocopy the test materials in any way under any circumstances.

After arriving at the school on the testing day, the field investigator should ask the school coordinator to give the school questionnaires to the relevant person at the school to complete. The completed school questionnaires should be returned to the field investigator on the day of the testing.

If within-school sampling is to be completed on the day of the assessment, the field investigator should request a list of all sections from the school coordinator and begin the procedures for section and student sampling (where applicable), as well as the allocation of test booklets as outlined in the field investigator’s manual using the field notes.

Once the within-school sampling has been completed and students have been assigned test booklets, the field investigator should be directed to the space where the assessment will take place and set up the assessment space and materials.

Once students are seated, the field investigator should ask the students to clear their desks except for any materials they are permitted to have for the test and also hand out any stationary (pencils, erasers, notepads) to students who require them. Test booklets should then be distributed to all students.

The testing session should begin and any late students should not be permitted to enter the testing room at this point. The field investigator should begin the testing session by following the instructions outlined in the field investigator’s manual, paying particular attention to any scripts that need to be read out to the students as they appear in the manual.
The testing session should include the following steps, which should be described in the field investigator's manual:

- introduce yourself and give students a brief overview of the purpose of the test using the pre-written script from the field investigator’s manual (no late students should be permitted to enter the testing room from this point)
- allow time for students to work through some practice questions, so they become familiar with how to answer the different types of questions
- show students how to complete the answer sheet – use the blackboard to demonstrate this and ensure all students understand what to do
- allow time for students to ask questions
- write the start time on the board (and use a clock or watch to time the session)
- monitor the students, ensuring that each student works independently throughout the testing session
- complete the student attendance sheet
- at the end of the allocated testing time, tell students to stop and close their test booklets – make sure everyone's book is closed
- collect the booklets and distribute the student questionnaires
- administer the student questionnaire using the same or similar procedures as outlined for the test booklets
- once students have completed the student questionnaire session, collect the questionnaires
- thank and then dismiss the students.

During and immediately after the completion of each testing session, the participation column of the student attendance sheet should be updated.

10.2.2 Determine if a follow-up session needs to be held

It is very important that a high level of student participation is achieved. If a significant number of students are absent on the testing day, a follow-up session may need to be scheduled for these absent students.

At the completion of testing, the field investigator should determine whether a follow-up session is needed based on the guidelines outlined in the field investigator manual. Should a follow-up session be required, the same procedures for test administration should be carried out.

10.2.3 Quality monitoring

To ensure the integrity and high standards of the within-school sampling and testing process, quality assurance procedures should be employed. Quality assurance monitoring of the within-school sampling and test administration involves the physical presence of a quality monitor at the testing site. The quality monitor should check the within-school sampling calculations and observe and report whether testing is taking place using standardised procedures under similar conditions during the main survey stage [4]. Having the within-school sampling calculations checked independently on the day ensures that if any errors are detected, these errors can be rectified before testing takes place.

A monitoring plan should be created for the main survey to take into consideration [4]:

- the number of unannounced schools the quality monitor will visit
- a method to select which schools will be monitored
- the creation of a list of schools for each quality monitor detailing the testing dates, address/location of school and the field investigator’s name.
10.3 Post implementation stage

10.3.1 Material collection and storage

A strategy for returning all test materials both used and unused should be planned for and communicated to the school coordinator and field investigator [3]. The field investigator should ensure that all test booklets (used and unused), answer sheets and student questionnaires, as well as the student attendance sheet and field notes, have been checked by the school coordinator. The field investigator should be responsible for the secure return of all these materials to the relevant authority immediately after testing.

The school coordinator should also make sure that the field investigator has correctly entered the participation information in the student attendance sheet. If a student booklet was unused (e.g., because the student was absent), the school coordinator should make sure that the field investigator has recorded this correctly on the student attendance sheet in the participation column.

Once all testing sessions and follow-up sessions have been completed and school coordinators have checked the student attendance sheet and field notes, the materials should then be returned to the relevant authority. The package should include:

- all completed student attendance sheets (there may be more than one depending on how many testing sessions were administered)
- a package of both the completed and the unused test booklets, sorted in order in which they are listed in the student attendance sheet
- a package of both the completed and the unused student questionnaires, sorted in order in which they are listed in the student attendance sheet
- the completed school questionnaire.

10.3.2 Receipt of materials from schools

Once all test materials and supporting documents have been received by the relevant authority, these materials should be carefully checked and steps should be undertaken to retrieve missing materials. The following checks could be incorporated:

- check that each participating school listed on the sampled school list has returned all materials
- check that the appropriate testing materials were received for every student listed in the student attendance sheet
- verify all identification codes on all test materials (e.g., all school codes are the same for all test materials received from the one school)
- check that the participation status recorded on the student attendance sheet matches the information on the test materials
- check that there is a field notes document for each school
- if within-school sampling was completed on the day of testing and the list of sections and/or students was on a separate sheet to the field notes, check to ensure that the original list or a photocopy of the list has been included
- check that all quality monitor checklists have been returned.
10.3.3 Quality assurance of within-school sampling process

In addition to the checks the quality monitor performs, the within-school sampling calculations should also be checked for each participating school to ensure that this process was carried out correctly. These checks should be performed at the district level. Clear instructions on how to carry out these checks should be created. Documentation of the findings should also be summarised and retained for reporting purposes and future studies.

The test materials are then ready for the next stage – the data entry, cleaning and verification (discussed in Chapter 11).
10.4 Summary

This chapter has outlined the activities and resources needed prior to, during and after the testing day. Standardised test administration procedures are vital to ensure that every student is assessed under the same conditions. The comparability of data across students and schools is directly affected by the extent to which standardisation is maintained. Quality assurance processes will identify areas of weakness and can provide confidence to the end users of the results that standardisation of the test conditions was achieved.

Clear documentation, procedures and team roles are essential components of the field operations. In particular, the documentation of the procedures within the school, such as the within-school sampling and the listing of student absences, are essential for subsequent steps in the assessment program workflow such as data cleaning (Chapter 11) and calculation of sample weights (Chapters 9 and 11).
### Checklist: Materials production and field operations

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th><strong>Developing the operation plan</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Has an operation plan containing a comprehensive list of activities, resources and timelines been created?</td>
</tr>
<tr>
<td>2. If individual operation plans are created, has one been created for each key role?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Developing the assessment materials</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Have all assessment materials including test booklets, questionnaires and answer sheets been developed?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Developing the administration materials</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Has a list of ALL administration manuals required for each stage of the project (field trial and main survey) been created?</td>
</tr>
<tr>
<td>5. Does each administration manual include a brief overview of the purpose of the test and the way the data will be used?</td>
</tr>
<tr>
<td>6. Does each administration manual include details of the roles and core responsibilities of the personnel for which the manual is intended?</td>
</tr>
<tr>
<td>7. Does each administration manual include step-by-step instructions for each core responsibility?</td>
</tr>
<tr>
<td>8. Does each administration manual include a description of the resources and preparation that are needed for each task?</td>
</tr>
<tr>
<td>9. Does each administration manual include a list of contact details of relevant authorities?</td>
</tr>
<tr>
<td>10. Have all procedural changes identified at the field trial stage been reflected in the main survey manuals?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Developing the field notes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Do the field notes allow for within-school sampling data to be recorded (either before or on the testing day)?</td>
</tr>
<tr>
<td>12. If the field notes were not used to record and perform the within-school sampling, has an additional note been included to ensure that this information is returned to the relevant authorities?</td>
</tr>
<tr>
<td>13. Has the field investigator been assigned the task of completing the field notes?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Developing the student attendance sheet</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>14. Does the student attendance sheet allow for the recording of student-level data, including attendance, booklet assignment and identification variables?</td>
</tr>
<tr>
<td>15. Has the field investigator been assigned the task of completing the student attendance sheet?</td>
</tr>
</tbody>
</table>
### Quality assurance of the materials

16. Have specific requirements relating to the quality of ALL print materials been created?  
17. Have ALL print materials been edited and reviewed prior to mass production?  
18. Has a printing company been sourced to undertake the production of the print materials?  
19. Has a printing plan been developed (including specifying print quantities, packing and collating instructions, distribution or delivery of materials and quality assurance)?

### Training

20. Have all personnel been identified and recruited (including field investigators and quality monitors)?  
21. Has a training plan been developed (including outlining the number of sessions, content and personnel involved)?  
22. Have all training materials been sent to key personnel in a timely manner?  
23. Have high-quality training sessions been delivered covering key field operations topics?

### Contacting schools

24. Has a plan been developed detailing when sampled schools should be contacted and when replacement schools of non-participating sampled schools be contacted?  
25. Has a letter been sent to school heads of sampled schools, requesting their schools’ participation?

### On the testing day

26. Does the field investigator have all of the necessary resources for the testing day?  
27. Has the school questionnaire been provided and the completed questionnaire collected?  
28. If the within-school sampling is being completed on the testing day, have these details been recorded in the field notes?  
29. Has the testing session been conducted following the procedures outlined in the field investigator’s manual?  
30. Has the student attendance sheet been completed?  
31. Have quality assurance monitoring visits occurred according to the monitoring plan?

### Material collection and storage

32. Has a strategy for returning all test materials been planned?  
33. Has the strategy for returning all test materials been communicated to the school coordinator and field investigator?  
34. Has a checklist detailing all returned materials been created and completed?

### Quality assurance of within-school process

35. Is someone responsible for checking all within-school sampling?  
36. Have the within-school sampling checks been documented and the findings summarised?
<table>
<thead>
<tr>
<th>Glossary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Administration materials</strong></td>
</tr>
<tr>
<td>Manuals relating to the administration of the tests (otherwise known</td>
</tr>
<tr>
<td>as field guidelines or field operations manuals) as well as important</td>
</tr>
<tr>
<td>supporting documents such as student attendance sheets (sometimes</td>
</tr>
<tr>
<td>referred to as student tracking forms) and field notes.</td>
</tr>
<tr>
<td><strong>Assessment materials</strong></td>
</tr>
<tr>
<td>Test booklets, questionnaires (eg student, school, teacher and/or</td>
</tr>
<tr>
<td>parent) as well as associated supporting documents such as answer or</td>
</tr>
<tr>
<td>optical character recognition/optical mark recognition (OCR/OMR) sheets.</td>
</tr>
<tr>
<td><strong>Field notes</strong></td>
</tr>
<tr>
<td>Document used to collect important school-level characteristics,</td>
</tr>
<tr>
<td>participation status, identification variables and within-school</td>
</tr>
<tr>
<td>sampling data for each participating school.</td>
</tr>
<tr>
<td><strong>Field trial</strong></td>
</tr>
<tr>
<td>Administration of items under test conditions, used to test the items'</td>
</tr>
<tr>
<td>validity and the administration procedures. Occurs before the main</td>
</tr>
<tr>
<td>survey with a sample of at least 200 students per item who are similar</td>
</tr>
<tr>
<td>to the target population.</td>
</tr>
<tr>
<td><strong>Main survey</strong></td>
</tr>
<tr>
<td>The final data collection stage where the data obtained is analysed</td>
</tr>
<tr>
<td>with the aim of making generalisations and inferences about the desired</td>
</tr>
<tr>
<td>target population.</td>
</tr>
<tr>
<td><strong>Non-participating school</strong></td>
</tr>
<tr>
<td>A sampled or replacement school that does not take part in the</td>
</tr>
<tr>
<td>assessment for some reason.</td>
</tr>
<tr>
<td><strong>OCR/OMR sheets</strong></td>
</tr>
<tr>
<td>Scannable forms using optical character recognition/optical mark</td>
</tr>
<tr>
<td>recognition software that are customised and used as score sheets.</td>
</tr>
<tr>
<td><strong>Operation plan</strong></td>
</tr>
<tr>
<td>A comprehensive list of activities and resources and the timelines to</td>
</tr>
<tr>
<td>which they need to be completed.</td>
</tr>
<tr>
<td><strong>Replacement school</strong></td>
</tr>
<tr>
<td>A school used to replace a sampled school should the sample school</td>
</tr>
<tr>
<td>be unable to participate. The replacement school closely matches the</td>
</tr>
<tr>
<td>sampled school on pre-defined criteria.</td>
</tr>
<tr>
<td><strong>Student attendance sheet</strong></td>
</tr>
<tr>
<td>Document used to collect student-level information such as</td>
</tr>
<tr>
<td>identification variables, test booklet assignment and participation</td>
</tr>
<tr>
<td>status for each test session.</td>
</tr>
</tbody>
</table>
References


Data management
Data management

This chapter aims to provide guidance on the best practices in data management while taking into account the unique contexts of large-scale assessments at the national and regional levels in India.

Although the general issues of data management are presented and discussed in this chapter, data management procedures are linked to the specific assessment project that they apply. As such, comprehensive details on the operational aspects of data management can be found in project-specific publications such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) technical reports [1, 2].

24 Although the issues discussed here are generally applicable to a wide range of assessment types, the scope of this chapter is mainly for large-scale assessments that have paper-based administration process.
11.1 Codebook design

The data codebook contains the information about all the variables in the data including their definitions, data type, values, validity parameters (e.g., number of characters/decimals allowed for valid values of each variable in the database), and codes for missing values. An electronic codebook is therefore a structural database which stores all the information given above but also includes the specifications for file structure, the coding schemes, the data verification rules and the quality standards for the datafile.

The design of the codebook is an important step in data management because it specifies how the raw data from the instrument are structured into the datafile. The analytical requirements dictate the structure and format of the datafile. For example, if the study design includes linear regression analysis, the data format of the outcome variable would need to be a continuous variable. As such, the design of the codebook should take into account the planned analyses so that the raw data are converted appropriately.

The design of the codebook should also take into account the allotment process for specific codes (e.g., codes for missing data) and should ensure that the format of the codes match the target variables. For example, if codes are assigned for schools, the number of digits for each school code is appropriate for the number of schools that need to be allocated unique codes. Similarly, the design should ensure that the codebook does not allow codes allocated for specific purposes (e.g., 9 as a code to indicate missing data) to be entered as variable values.

The codebook format is dependent on the data management software that will be used and it can range from common (e.g., Excel or Access) to specialised (e.g., PostgreSQL, WinDEM [3], KeyQuest [1]).

Regardless of format or design, every codebook must contain the following:

• information about the variables in the dataset, including units of variables if applicable (e.g., variable name = STU_AGE, variable label = student age, unit = years)
• information about the summary methods that were implemented (e.g., which variables were combined, aggregated, or converted into composite variables)
• information about the study design (e.g., whether it is a longitudinal study, which variables refer to each of the various time points).

---

25WinDEM is used in the Trends in International Mathematics and Science Study (TIMSS).
26KeyQuest is used in the Programme for International Student Assessment (PISA).
11.2 Data management software

Large-scale data collection is a complex process that can involve a large team and numerous cycles of data collection over an extended period of time. The use of data management software can substantially streamline and increase the efficiency of this process. The choice of data management software may depend on the type of data collection and entry. The three main types are outlined below.

11.2.1 Manual entry

The most basic option for data collection is through paper forms that need to be manually handled. This option involves a large amount of human resources and provides the least amount of data reliability. Nevertheless, the weakness of manual data entry can be mitigated through the use of data management software. WinDEM and KeyQuest are two examples of data entry management software that can systematise manual data entry. The main benefits of this software include data validation, automatic error checking, data verification, file checking, and flexible reporting for data audit.

11.2.2 Optical character recognition/optical mark recognition (OCR/OMR)

A second option is to revise the data collection from manual forms to scannable forms using optical character recognition/optical mark recognition (OCR/OMR) software. This option reduces the bulk of the materials because the questionnaires become reusable (only the OMR scan forms need to be printed for each student per test) and it greatly reduces the time requirement for converting the data into digital formats by automating data entry. The OCR/OMR software has a custom data management component that allows consolidation and structuring of the scanned data from custom score sheets.

There are several types of OCR/OMR platforms. The traditional platform (eg Scantron) is a package that bundles the hardware (scanner machines) and the software. This type of platform usually needs dedicated scannable forms. Newer platforms are device-independent, such that the hardware and software do not need to be purchased from a single source. A range of device-independent commercial software (eg Remark) as well as open-source software (eg TCExam, SQS, FormScanner) is available. In both types of platforms, the score sheets need to be designed, either using a template or a custom design, to accommodate OCR/OMR scanning.

It is recommended that device-independent software be used for data capture so that there is flexibility regarding the choice of scanning hardware. Minimum specifications for the scanning hardware need to be mid-level to high speed (faster than 60ppm) and high-capacity (greater than 10 000 scans per day) for assessments that involve more than 100 schools. Scan forms from the test sites normally need to be transported and centrally processed. The data are then stored in the native format in secure drives at the centre. Depending on the OCR/OMR software, secondary data management software may be needed. Data in native format will often need to be converted to other formats for statistical analysis. This conversion may be done by a bundled software package or processed externally.

Quality assurance is important in OCR/OMR production. Quality assurance involves checking the printed forms, conducting random checks of the scanning output, and double scanning (see Section 11.4).
11.2.3 Digital entry

The third option is to use a completely digital process for data collection and scoring. This requires hand-held mobile devices for data capture and entry. Small digital tablets can be used as a medium for questionnaire delivery, as well as a data entry platform where the field investigator directly inputs data in digital format. This option is the most efficient in terms of logistics because no forms need to be printed and physically transported. The data entry and scoring will also be seamless due to completely digital transfer and processing. However, this option entails considerable costs, in the development of the software, development of the items and the purchase of the devices. In addition, this option requires the highest technical know-how for the field investigators, a requirement that may be very difficult to satisfy given local contexts and availability of personnel.

Data from these devices may be consolidated physically (e.g., through physically collecting flash drives and compiling the data centrally) or digitally (e.g., remote transmission of data). Similar to OCR/OMR data collection, there are device-dependent and device-independent data management software options. Device-independent software can take raw data and process them independent of the data collection devices. These device-independent software options are recommended because they often are more flexible and have wider range of data management features.
11.3 Physical materials handling

It is important to design the protocols for physical materials handling carefully because physical handling is the most logistically challenging. Physical handling of materials also increases the chances of errors to occur.

Systematic procedures, including the replacement of physical handling with automated process wherever possible, and robust records-keeping can help minimise data loss and reduce human error during the physical handling of assessment materials.
11.4 Data entry protocols

11.4.1 Data preparation

Large and complex studies with very demanding standards for data quality require an extensive set of interrelated data checking and cleaning procedures. To ensure that all procedures are conducted in the correct sequence, that no special requirements are overlooked, and that the cleaning process is implemented independently of the people in charge, the following steps should be undertaken:

- Before their use with real data, all data-cleaning programs/syntax should be thoroughly tested using simulated/dummy datasets containing all possible problems and inconsistencies.
- Additional variables like language code and uniquelID may need to be created for each dataset.
- All incoming data need to be maintained in a specific database. The date of arrival is to be recorded, along with the datafile name. Filenames of all components (datafiles, codebooks, changelogs, syntax, etc) should be consistent throughout.
- The cleaning should be organised following strict rules laid out in a data cleaning manual.
- All corrections to datafiles should be listed in a cleaning report.
- Occasionally, it will be necessary to make changes to datafiles. Every such ‘manual’ correction must be logged for documentation.
- Once the data cleaning is completed for a subject, all cleaning steps should be repeated from the beginning to detect any problems that might have been inadvertently introduced during the cleaning process.
- It is important that robust backup procedures are implemented (see Section 11.8) to make sure that the original dataset is retained.

11.4.2 Missing data recoding protocols

It is important that the recoding protocols have clear rules on the treatment of missing data. When there is no response on a data collection instrument for a particular question, this may have different reasons. For coding and scoring purposes, it is essential to distinguish between data that are missing by design because the respondent was not administered a particular data collection instrument, and data that are omitted because the respondent did not respond to a question, even though he or she had the opportunity to do so. This distinction is important because different cases of missing data will be treated differently in data analyses.

All missing items are considered incorrect for scoring in whenever the test has been administered. This includes situations where students have not finished the test as well as situations where the omission was due to a triggered stop rule. Stop rule involves a test design assumption that the subsequent items (where the stop rule applies) are no longer administered because the student will be unable to do them, and therefore administering these items will only cause undue stress and extend the administration time.

Omitted items may be considered missing data when the items were not administered for reasons beyond the control of the test-taker (eg when the test is interrupted due to an emergency).
11.4.3 Data entry monitoring

It is good practice to monitor the data entry operators on a daily basis. Each of the data entry operators should be assigned a unique code. A variable should be created in the database so that a record can be kept about who has entered the case. A daily sample check (around 3 to 5% of data) for each of the data entry operators should be carried out to ensure valid entries into the database. Manual or automated checks can be done on the selected samples as part of the data entry audit. The data manager should conduct the monitoring and audit of the data entry operators. International best practice is to tolerate no more than 2% error in data entry.

Double entry or double data-punching may be implemented to complement the monitoring process. Double entry procedures have logistical and financial implications so they need to be carefully planned. It is recommended that at least 10% of the data be entered twice to assess the quality of the data entry. This double entry check should be applied only to the manually inputted data (this includes manually scanned OCR/OMR forms). It is on the assumption that the same random data entry error is unlikely to appear simultaneously and therefore, most data entry errors would be identified as a discrepancy between two parallel sets of data entered by different data entry operators. The same principle can be applied in OCR/OMR procedures to detect transcription errors by implementing double scanning. This check becomes more effective if the double scanning is done on different machines. A thorough manual check needs to be done on the scanned forms whenever discrepancies are identified between the two scans.

The data manager may increase the proportion of the data to be double punched depending on the available resources and the perceived reliability of the data entry operators.
11.5 Data management training

To prepare for the data entry, a national assessment centre should designate a data manager [4]. The data manager is the member of the national team who will be responsible for data entry and the training of data entry operators (personnel who enter the data). Data managers will also supervise the data cleaning, verification and validation stages.

The data manager preferably should be a staff member dedicated to data management tasks. If this is not possible, the project manager could also assume this role. However, this may lead to problems due to competing duties that project managers may have at the time of data entry. Data managers should be thorough and have some experience in dealing with managing larger sets of data.

While data managers can be trained by external consultants, it is usually the responsibility of each project centre to employ and train their data entry operators. These activities will require additional staffing and resources, and therefore it is recommended to commence planning for these activities well in advance of the testing period, in particular if national centres foresee problems with the recruitment of suitable data entry staff.

Training should focus on the following aspects of data collection and management:

- **Field operations**: This part of the project includes packaging and shipping materials, test administration, receipt of materials at the national centre after testing, and coding of student responses. Those responsible for the logistical aspects of the assessment will need proper training in this area.

- **Data entry**: Training on the use of data entry software is needed for the data entry operators and the data manager. The software may or may not be preconfigured with the structure and format of the standard instruments: cognitive test booklets, background and contextual questionnaires, and a student tracking instrument. As such, training will need to be extensive for the data managers to be able to fully customise the software according to the data management needs. The data manager can be trained by consultants and transfer the necessary skills for data entry to the data operators.

- **Post data collection processing**: This stage involves data cleaning, verification and validation. Training in each of these areas is needed for the data manager and/or data administrator. The training will focus extensively on data security, database management, data transfer and quality control. The process for data cleaning, verification and validation is discussed in more detail in the next sections.
Data cleaning, verification and validation

Data cleaning is one of the major components of the data quality. Protocols for data cleaning need to be specified in a project manual. International best practices in data cleaning, verification and validation can be adopted by referring to the published data management manuals [eg 1, 2].

11.6.1 Preparatory activities

Manual scrutiny of the collected data is typically the first preparatory activity in the data cleaning process. It is recommended that the data cleaning process starts at the earliest stages of data collection so that errors are detected as close to the source as possible. This means that test supervisors need to be briefed on quality control processes and standards (see Chapter 10 for more details on the field operations quality control).

Involving relevant staff and putting in place data cleaning procedures at all stages of data collection also enables a decentralised system for the preparatory activities of data management. In the context of India, a decentralised preparatory stage spreads the burden of manual scrutiny and physical materials handling, which has the advantage of increasing workflow efficiency and reducing central points of failure.

11.6.2 Data cleaning for multi-stage data collection

For multiple-stage data collection, data cleaning must be done each time the data are moved up one level. For example, when collecting data from schools which are then consolidated by district, data cleaning must be done at the district level. Then when these district-level data files are consolidated again at a regional level, another round of data cleaning must be conducted (see Exhibit 33). During the process of data cleaning, each upper level will send cleaning reports containing the results of the checking procedures to the lower levels, and will ask the lower levels to clarify any inconsistencies in their database. In the questionnaires, for example, such inconsistencies might include cases where the number of qualified teachers in a school exceeds the total number of teachers. The upper level datasets are then updated according to the information provided by the lower level.

Exhibit 33: Data cleaning in multi-stage data collection
When data are submitted by local centres, the first step is to check the consistency of the database structure with the national database structure. During this process of data cleaning, queries, if necessary, are sent to local centres. The local datasets are updated, to be consistent with information provided by the local centres. For example, if a variable has been added to a questionnaire at a district level, the questionnaire adaptation sheet will be checked to find out whether this district-level variable requires recoding into a corresponding national-level variable, or has to be set aside as a district-level option. Once all deviations have been checked, the assessment team will apply necessary recodes for the submitted data to ensure that the local data match the national database structure.

11.6.3 Violations in the keys

In database terms, the ‘key’ is a combination of identifier codes which usually include strata, school and student identification. Each record in a database will have a unique key. For example, if a student identification code is part of the key, it should be unique within the school. In conjunction, unique school codes need to be included as part of the key to ensure that each complete key is unique.

Key violations can be introduced during data entry if a data entry operator enters the same combination twice or enters a combination that does not exist (e.g., a school identification code that was not listed for the sampled schools). Key violations almost always occur at the lowest level of data collection (i.e., bottom level of Exhibit 33) and the further up an error is detected, the harder it is to resolve. There needs to be clear procedures for checking key violations and a system for sending cleaning reports to the local centres to clarify identification codes. Further checks are done in the data verification and validation stages.

11.6.4 Duplicates

One of the most common data entry errors is having duplicate records. Duplication may occur for an entire respondent, or it can occur to a subset of responses or variables. Performing a duplicate search will depend on the data management software but even common programs such as Excel have some functionality for detection of duplicates. Some software programs, such as WinDEM, have very comprehensive and thorough duplicate checks that can handle multiple identification variables.

11.6.5 Data consistency

Data consistency within the datafile involves checking that values are valid, within range and consistent. For example, data for month/year of birth must be formatted consistently and should be within range. Frequency checks are useful in detecting out-of-range values as well as checking the number of invalid values (e.g., missing values that are not properly coded as missing).

Frequency checks are also useful in detecting values that are technically valid and within range but are incorrectly entered. For example, when entering month data, an incorrect double key-press of 1 for January would result in a valid value of 11 (for November). This could not be flagged by validation rules but can be checked by comparing frequencies from secondary sources.
Missing data on key variables, such as respondent identification codes, are especially important to check. Consistency in format (e.g., number of decimal values) is also important, especially for variables that are components of further calculations. Most software programs, even Excel, have data validation tools which can simplify the detection process for inconsistent data.

Merging multiple files can introduce errors if checks are not done properly. Records can be dropped and human error can cause some files to be excluded during the merging process. A simple check of the batch totals can detect missing records. It is also important to conduct duplicate checks in case one or more files have been merged more than once. These duplication issues are particularly prone in situations where several people are handling the merging process.

11.6.6 Linking variables

Linking variables are variables that are common across multiple datafiles (e.g., school code for student response, teacher questionnaire and school head questionnaire datafiles). Linking variables are used to link or merge several datafiles. Consistency between datafiles must be checked, especially for linking variables. For example, linking identification codes must have exactly the same data format. Even minor inconsistencies (e.g., hidden spaces) may result in merge failure when linking between datafiles. In SPSS it is important to do a structure check to ensure that the type, format, and categories are consistent for all linking variables. In WinDEM, the codebooks must be checked to ensure that the linking variables are consistent.

11.6.7 Logic checks

For some variables, the valid responses may depend on responses to other variables. For example, certain questions may only be applicable for students undertaking particular subjects, or the range of values for an item may be different for younger and older children. Variables that are components of calculation formulas should also have inputs that do not yield a logical error (e.g., an input value should not be 0 when it will divide some other number). Logic checks need to be performed for these variables using script-based validation checks. Excel and SPSS are common software programs that can perform these types of logic checks.

11.6.8 Resolving issues

Regular audit during the data collection and entry helps to minimise data errors. Regular audit also allows errors to be detected earlier in the process, making the issues easier to resolve. For example, if a missing date of birth is detected while still in the field, it is easier to obtain the correct information. The data manager should resolve the issues whenever they arise by checking the raw data and linked datasets. Whenever an issue is resolved, backups must be done before changes are made.
11.6.9 External checks

Some discrepancies may need checks using external sources of data to resolve. External sources include school records, census data, public records and other pre-existing databases that are separate from the current data collection process. For example, basic demographic data (e.g., students’ sex and age) may need to be compared between the student list received from schools and data captured from the responses in the student questionnaire. During the data processing stage, cleaning and/or verification reports with discrepancies are sent to the local centres for clarifications. The national datasets are then updated according to the information provided by the local centres.

11.6.10 Procedures for unresolved issues

If unresolved inconsistencies still remain after all above procedures have been implemented, the following general rules apply:

- Unresolved inconsistencies regarding student and school identification will lead to the deletion of the record from the database.
- The data for any cognitive item affected by an unresolved systematic error will be replaced by the Not Valid code. For instance, for mistranslation or misprinted items, the data for these items will be recoded as Not Valid and will not be used in future analyses.
- If a variable from the background questionnaire has been deleted between cycles of data collection, it will be replaced with the Not Administered code for the current cycle.
- If a cycle of data collection includes a modified a variable such that it cannot be matched into the previous cycles, it will be renamed (and treated as a new variable) and the data for the original variable will be replaced with the Not Valid code.
- All data for added or modified background variables will be set aside in a separate file for different provinces to use for their own purposes.
11.7 Database construction

Project databases need to be designed for the user. Depending on the project mandate, there may be external users that may need to access the data and therefore specific databases need to be constructed to fit the users’ needs.

Two types of external users are common in educational research settings: the general public and the institutional user. The availability of the project database to the general public is often a requirement in projects that have a mandate for transparency. Institutional users typically require access to the database in order to conduct secondary analyses on the data. As these two types of users often have different needs, the design of the database may need to take into account the target user.

It is important to note that even when the aim is transparency, privacy rules still apply. As such, it is important to ensure that private and confidential data are not accessible in the public databases.
11.8 Backup procedures

There should be regular and systematic backup protocols. These backup procedures need to be done for both raw and converted data. Raw data need to be backed up before exporting into another format. Data should also be backed up whenever datafiles are consolidated. The codebooks should also be backed up before any structural changes are made. Backups should follow a systematic file-naming protocol. The filenames should at least include the date, backup type (eg full, incremental), stage of data processing and source of data.

A change log needs to be kept and should include the following information about the file’s version history:

- what was changed
- when the change was made
- who made the change
- why the change was made.

There should be a regular scheduling of backups; preferably at least daily archives should be kept. There are archival software programs that can systematically manage and automate this process. It is important that backups are kept in a separate data storage device from the originals, and preferably in a physically separate location. Cloud storage is very useful, although data security, ownership and ethical issues need to be considered carefully. Due diligence must be conducted before any large-scale data transfer to the cloud is implemented.
11.9 Database security

Database security needs to be designed into the data management software. Access rights need to be specified and implemented in the software to allow different types of users to access and manage the data. The most common types of users and their respective access rights are:

- Data entry operators – these users can enter the data and run data verification procedures such as checking for duplicates and running validation checks. Data entry operators should not be able to import/export data, delete records, create and edit codebooks, or change the software settings.
- Data managers – these users can import/export data, create/edit codebooks, delete or change records, but they cannot add or delete users nor set or change the user rights. The number of data managers depends on the scope of the assessment project. Smaller projects might have a single data manager while large national projects may have data managers for each hierarchical level of the project (i.e., region, province and district).
- Data administrator – the data administrator has full access rights. In smaller assessment projects, the role of the data administrator can be incorporated into the data manager’s role.

The access rights must be defined for all forms of datafiles, whether they are raw, processed or archived.
Data transfer protocols

Data must be transferred and/or shared in a self-contained structure. The general form of any self-contained set contains the following:

- the contents (raw data or processed data)
- documentation about the contents: source, structure and semantics
- instructions on how the data was manipulated (if applicable, so that it can be reverted into a previous form if necessary)
- instructions on how the contents can be viewed, manipulated and linked to other datasets.

Transferring these self-contained sets must be done by the data manager and systematic records must be kept for each transfer.
11.11 Quality control

Quality control measures for data management are embedded in each of the stages described in this section (data entry, data cleaning, verification and validation). These measures need to be documented in a single quality control manual and reviewed periodically. The data manager is responsible for developing this manual and disseminating this to all staff who handle the data for continuous feedback. It is important that lessons from the field are incorporated into the quality control review cycle. In a continuous or rolling assessment process, the quality control manual needs to be reviewed and finalised before the data collection phase for the next cycle begins.
Summary

This chapter has outlined the data management process for a large-scale assessment program. It focused on both broad issues, such as data cleaning, as well as more specific and operational issues, such as data entry protocols. This chapter also discussed supporting components such as human resources and training requirements that, although not exclusive to data management, have important roles in the data management process.

Data management is a critical component of any assessment project. The quality of results from statistical analyses is contingent on the quality of the data, which in turn is contingent on the quality of the data management process. This chapter discussed the key issues to consider in data management, covering major data management steps starting from the design of data collection instruments all the way to data transfer and security issues.
### Checklist: Data management

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Codebook design</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Has a codebook for the data been designed according to the needs of the research design and data collection requirements?</td>
</tr>
<tr>
<td>2. Does the codebook contain the important components outlined in the section?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data management software</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Is the data management software selected appropriate for the context of the program?</td>
</tr>
<tr>
<td>4. Is the software appropriate for the data collection requirements and is it compatible with the codebook structure?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data entry protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Have protocols for data entry been set and properly documented?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. Have the roles for all staff involved in data management been clearly defined?</td>
</tr>
<tr>
<td>7. Is there a training program for all staff involved in data management?</td>
</tr>
<tr>
<td>8. Does the training program cover all important aspects of data management?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data cleaning, verification and validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Have protocols for data cleaning, verification and validation been set?</td>
</tr>
<tr>
<td>10. Are the procedures for multi-stage data processing appropriate for the contexts of the program?</td>
</tr>
<tr>
<td>11. Have ways to improve the multi-stage data processing been considered?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database construction, backup, security and transfer protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>12. Have protocols for database construction, backup and security been set?</td>
</tr>
<tr>
<td>13. Have the user types and access rights been established?</td>
</tr>
<tr>
<td>14. Have quality control standards and procedures been set?</td>
</tr>
<tr>
<td><strong>Glossary</strong></td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td><strong>Access rights</strong></td>
</tr>
<tr>
<td><strong>Archival software</strong></td>
</tr>
<tr>
<td><strong>Change log</strong></td>
</tr>
<tr>
<td><strong>Cloud storage</strong></td>
</tr>
<tr>
<td><strong>Converted data</strong></td>
</tr>
<tr>
<td><strong>Double entry or double data-punching</strong></td>
</tr>
<tr>
<td><strong>Raw data</strong></td>
</tr>
</tbody>
</table>
References


Scaling methodology
Scaling methodology

Assessment results are better interpreted and therefore more useful if they are reported on a scale sharing a common unit for measuring and interpreting results.

Scaling is the process of constructing a score scale that associates numbers or other ordered indicators with the performance of examinees [1]. The process of scaling produces a score scale, and the resulting scores are referred to as scale scores. Petersen, Kolen and Hoover [2] stated that ‘the main purpose of scaling is to aid users in interpreting test results’ (p. 222).

The choice of scaling methodology is fundamental to the overall assessment design and an important technical consideration for proper interpretation of data. This chapter will look at Classical Test Theory (CTT) for describing assessment results and outline the benefits of using Item Response Theory (IRT) scaling methods instead. Applying advanced scaling methods is a highly technical activity and has implications for obtaining or developing appropriate expertise. These methods will be described and compared and recommendations for best practice will be made.

There are many different methods to constructing and reporting scales. Some familiar scales include the A to F grades reported in some schools; or the 0 to 100% reported on examinations; and the various National Achievement Surveys (NAS) in India report results on scales that have a mean of 250 and a standard deviation of 50. Current best practice in large-scale surveys is to report on scales that are constructed through the application of IRT.

Finally, this chapter will sketch how the results can be made more interpretable, by creating qualitative descriptions of the quantitative scales that are recommended.
12.1 Scaling models

Scaling models are driven by different theories. This section briefly introduces CTT and IRT, and the scaling models based on these theories. It will also explore why IRT models are more suitable for scaling large-scale assessments.

12.1.1 Classical Test Theory models

CTT [3] models are based on true-score theory. CTT focuses on estimating each student’s ‘true score’ and making inferences about his or her likely score on a test. The test score of CTT ranges from 0 to the maximum score on a test. If a mathematics test is given to students, CTT estimates the likely students’ scores on this particular mathematics test. Although the item difficulties and student results can be viewed alongside each other to aid interpretation, CTT is limited to comparing scores on the same test. There is little scope for generalising skills of students at specific ability levels. Comparing student performance over time is not possible unless the same tests are used each time. In summary, CTT is suitable to use only for analysing and ranking students’ scores from the same test.

Under CTT, when raw percentages are used to measure students’ abilities and item difficulties, it is not immediately obvious how one can link student scores to item difficulties. For example, Exhibit 34 shows two scales: one for item difficulty, defined by the percentage of students that answered the item correctly, and the other for student ability. The item difficulty scale on the left shows 20% of the students answered the word problem item correctly (Item 1). In contrast, 80% of the students answered the single-digit addition item correctly (Item 2). This indicates that Item 2 is relatively easier than Item 1.

On the other hand, the student ability scale shows four students who answered 80%, 70%, 40% and 20% correctly for the same test. The percentages on the two scales are not easily matched. Inference of student learning cannot be made because the proportions of items by type are not known. For example, it may be the case that 80% of the items are single-digit addition items, so that the students who obtained 80% correct on the test cannot perform tasks which are more difficult than single-digit addition.

Even if there is information on the composition of the test in terms of the number of items for each type of problem, it is still difficult to match student scores with items. This is because there is no simple and direct relationship between student scores and item difficulty, and the underlying skills and descriptions for each student score will need to be studied and reported separately. As a consequence, when a different test is administered, a new set of ability descriptions will need to be developed.
12.1.2 Item Response Theory models

Item Response Theory models focus on estimating each student’s ability and making inferences about each student’s ability level on the construct (i.e., a latent trait such as intelligence, motivation or maths ability) that is being tested. According to Carroll [4], ability is defined as the ‘possible variations over individuals in their threshold levels of difficulty in successfully performing some defined class of tasks’ (p. 4).

Because constructs are latent, they must be indirectly measured through related observables. A score on a questionnaire or test corresponds to the set of responses on the observed variables (the items), which in turn are indicators of the latent construct (Exhibit 35).

Exhibit 35:
Relationship between the latent and observed variables under Item Response Theory

Unlike CTT, where student ability is described within the boundaries of 0% and 100% correct on a test, a latent trait is measured on an infinite continuum, where the measurement unit is denoted as a logit. If a mathematics test is given to students, the IRT approach would try to estimate each student’s level on the latent trait of mathematics. The logit defines distances between differences in scores which can be easily interpreted. It also can link item scores to person scores. IRT offers more capacity than CTT for linking different tests and providing substantive interpretations to scores on a test. It helps in placing different tests on the same scale for comparison in time.
IRT models are probabilistic. They give the probability of a successful response to an item by a person. Under IRT, the item difficulty for an item is identified as the ability level with a 0.5 probability of success on the item. In the example given in Exhibit 36, the red line shows the probability of success on an item at each ability level:

- For a high-performing student (ذكر), the probability of success on this item is close to 1.
- For a low-performing student (ذكر), the probability of success on this item is close to zero.
- For a student of average ability (ذكر), the probability of success is 0.5.

The ability level of the average person ($\delta$) is also the item difficulty of this item. The notion of an item difficulty is defined as the ‘average’ difficulty of the item. Therefore, the item would be very easy for a very able person, and difficult for a low-ability person. A person who has the same ability level as the item difficulty ($\delta$) would have a 50% chance of being successful on the item.

**Exhibit 36:**
Probability of success

With the item difficulty and person ability, one can compute the probability of success for each person on each item using IRT models such as the Rasch model [5] and Birnbaum models [6].
12.1.2.1 Rasch (one-parameter logistic) model

The Rasch model is a latent trait model based on IRT. The Rasch model links the latent trait (i.e., student ability) with a single item parameter (the item difficulty) and as such, it is also known as the one-parameter logistic (1PL) model.

One important feature of the Rasch model [5] is to create a continuum on which both student performance and item difficulty are located with a probabilistic function linking these two components (Exhibit 36). In the case of dichotomous items, such as multiple-choice items or correct/incorrect items, the Rasch model predicts the probability of selecting a correct response (value of one) instead of an incorrect response (value of zero). It is modelled as:

\[
P_{in} (\theta_n, \delta_i) = \frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}}
\]

where \( P_{in} (\theta_n) \) is the probability of person \( n \) scoring 1 on item \( i \), \( \theta_n \) is the estimated ability of person \( n \) and \( \delta_i \) is the estimated location of item \( i \). For each item, item responses are modelled as a function of the latent trait \( \theta_n \).

For items with more than two \((k)\) categories such as Likert-type items, the more general Rasch partial credit model [7, 8] can be applied. It takes the form of:

\[
P_{xi}(\theta_n, \delta_i) = \frac{\exp \left( \sum_{t=0}^{h} \left( \theta_n - \delta_i + \tau_{ik} \right) \right)}{\sum_{k=0}^{h} \sum_{t=0}^{h} \exp \left( \theta_n - \delta_i + \tau_{ik} \right)} \quad x_i = 0, 1, \ldots, m_i
\]

where \( P_{xi}(\theta) \) denotes the probability of person \( n \) scoring \( x \) on item \( i \), \( \theta_n \) denotes the person’s ability, the item parameter \( \delta_i \) gives the location of the item on the latent continuum and \( \tau_{ik} \) denotes an additional step parameter for each step \( k \) between adjacent categories.

The analysis of item characteristics and the estimation of model parameters can be carried out using IRT estimation software (e.g., ACER ConQuest [9]).

12.1.2.2 Birnbaum (two- and three-parameter logistic) models

Birnbaum models [6] are also latent trait models based on IRT. The two-parameter logistic (2PL) model adds an item-specific discrimination parameter. The three-parameter logistic (3PL) model in addition includes a guessing parameter for each item. The general form of the Birnbaum models is:

\[
P_{in} (\theta_n, \delta_i, a_i) = c_i + (1-c_i) \frac{e^{a_i (\theta_n - \delta_i)}}{1 + e^{a_i (\theta_n - \delta_i)}}
\]

where \( a_i \) is the discrimination parameter and \( c_i \) is the guessing parameter for item \( i \). For the 2PL model, \( c_i \) is set at 0.
12.1.2.3 Comparison between the IRT models

A thorough comparison between different IRT models is beyond the scope of this handbook. There are however numerous references that discuss and compare these models in detail [eg 10, 11]. In brief, the 1PL (Rasch) model has a stricter definition of what constitutes measurement. Therefore, the measures derived from the model have stronger construct validity, which is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not ‘operationally defined’ [12]. Fitting 2PL or 3PL models will improve the model fit (see Section 12.4.2.1) but such models are less well-suited to construct validation.

Also, the 1PL model is more stable over time, which is good for measuring trends. This is because it is the simplest model, with only one parameter to vary on the item side. The more complicated the model, the more opportunity for the parameters (and their interaction) to vary and hence the measure of trends can become more problematic. Another advantage of the Rasch model is that item fit can be more strictly evaluated and a decision can be made about the extent to which an item is biased with respect to language, cultural background, gender, or any other factor.

The 1PL is the most preferred model for large-scale assessments, and as such, the rest of this chapter is focused on providing a general overview of scaling procedures using this model.
12.2 Methods for linking design

In educational measurement, it is commonly of interest to place the results of two or more tests on a common scale. The process of placing two or more tests on a single scale is called linking. Linking can also be applied for different booklets or test forms of the same instrument. In general, linking can be applied to a set of items, whether this set constitute a test or a test form. This process enables the difficulty between different tests and/or item banks to be compared.

The main advantage of linking is that it allows for instruments to measure a broader scale of the construct without being too lengthy, because subsets of items can be given to students. Linking also allows for the construction of different tests, which are on the same scale and measure the same construct, to span across levels (known as vertical linking or vertical scaling) and time/cycles (known as horizontal linking or test equating, see Section 12.3). Below are four types of designs [13] for linking different tests:

- **Single-group design**: This design links two tests by giving each test to the same group of students. This is a simple design; however, it can introduce student fatigue due to long test duration. In addition, the first test can be seen as a practice for the second test, which can influence the parameter estimates and the linking results.

- **Equivalent-groups design**: This design links two tests by giving the tests to equivalent but not identical groups of students whom are randomly chosen. This design has the benefit of avoiding the fatigue and practice issues. However, an equivalent-groups design can be costly and impractical to implement if random student sampling is used for large-scale analysis.

- **Common-person design**: This design links two tests by giving each test to two groups of students, where there is a common group of students taking both tests. Although this design allows linking two different tests, the group of students taking both tests would have the same fatigue and practice issues as the single-group design.

- **Anchor-test design**: This design links two tests by having a set of common items in each test. The two tests are administered to two different groups of students. If the set of common items is properly selected, this can avoid the issues of the other three designs discussed above. As such, the anchor-test design is frequently used in educational measurement. It is also desirable to overcome other test design issues, for example, through using the complete rotation of subtests to construct test booklets to avoid item positioning effects, where students may perform less well on items towards the end of the test (see Chapter 8 for further information on test design). The anchor-test design is also known as the common-item non-equivalent-groups design.
**12.3 Test equating**

Test equating is a form of linking that is often used for tests that were developed separately (but were designed to measure the same construct at the same level) or for alternate forms of the same test. Kolen and Brennan [1] define equating as a ‘statistical process that is used to adjust scores on test forms so that the scores on the forms can be used interchangeably’ (p. 2) and the process is meant to ‘adjust for differences in difficulty, not for differences in content’ (p.3). After the items of the two tests are assessed and the anchor items are finalised, equating can be carried out to link the two tests. Another statistical process that is related to equating is vertical scaling (also known, but less commonly, as vertical equating), which links two tests that differ in content.

It is important to note that equating is only as good as the quality and number of common elements. Some common equating methods are outlined below. Each can be suitable for vertical (eg different year levels) and/or horizontal (eg different cycles) equating.

- **Joint calibrations**: This method is used to combine data into one dataset and calibrate all items together. Joint calibrations must be done through the item linkages such as the final set of common items. The linkages can be ‘indirect’. For example, in a situation with three test forms, test 1 is directly linked to test 2, and test 2 is linked to test 3. In this case, test 1 is not directly linked to test 3 but is indirectly linked. It is useful if the datasets are equivalent, for example, through the use of different rotated test forms.

- **Anchoring**: Item parameters for link items in the new dataset are ‘anchored’ on the item parameters from the old dataset. These item parameters are not re-estimated for the new dataset. If a scale has been clearly established using a large dataset and some smaller datasets need to be calibrated on to the main scale, then the anchoring method can be considered.

- **Constant shift**: This method computes the difference of the item parameter average of the common item set between the two calibrations. The difference is then applied as a constant shift to all item parameters in the new dataset.

In any equating method, the sample of selected link items influences the estimated shift. It means that the resulting shift could be slightly different if an alternative set of link items is chosen. Therefore, there is an uncertainty associated with the equating, which is referred to as linking or equating error. This error should be taken into account when making comparisons between the results from different data collections across time.
12.4 Calibration

Scale calibration is the process of estimating the parameters of the model (e.g., item difficulty in Rasch) and placing these parameters on a uniform scale [1]. Calibration is done after the test data are received and cleaned. The calibration process is carried out to assess item behaviour. This section will discuss treatment of missing data, item review and assessment of common items.

12.4.1 Missing data

Missing student responses that are likely to be due to test length are treated as not reached items. Not reached items are defined as all consecutive missing values at the end of the test except the first missing value of the missing series, which is coded as embedded missing, like other items that are presented to the student but to which the student does not respond (e.g., if the student response on a 10-item test is as follows: 1011010000, response to item 7 is coded as embedded missing, while items 8 to 10 are considered not reached). Not reached items are omitted from the calibration of item parameters but treated as incorrect for the scaling of student responses. All other missing responses are included as incorrect responses for the calibration of items. Responses marked as not administered, such as items with printing errors or items in rotated designs that are not given to some students, are omitted in the calibration of items as well as from the scaling of student responses.

12.4.2 Item review

As described above, the Rasch model, through a probabilistic function, builds a relative continuum on which each item’s difficulty is located. The relative difficulty of an item results from the comparison of that item with all other items in the same test. The relative item difficulties are independent of the student abilities. During item review, one must carefully study the item statistics and fit to determine if an item is retained for the final scaling of student responses.

To produce unbiased item statistics for item review, it is recommended that for large-scale surveys each sampled group contributes to the item statistics equally. Otherwise, the item statistics could be dominated by states that are particularly large compared to the others (e.g., in India). To achieve equivalent contributions from different-sized groups, each group can be weighted to a uniform value (this is known as ‘senate weighting’) or an equal-sized subsample of each of the groups can be drawn for inclusion in the calibration (a calibration sample). Sampling and weighting are discussed in more detail in Chapter 9.

12.4.2.1 Model fit

The model fit of cognitive test items should first be assessed using a range of item statistics produced from the preliminary item calibration. The weighted mean-square statistic (infit) is a residual based fit statistic. It is used as a global indicator of item fit. Infit statistics are required to review for both item and step parameters. In addition, item characteristic curves can be also utilised to review item fit. Item characteristic curves provide a graphical representation of item fit across the range of student abilities for each dichotomous or partial credit item.
A useful tool for test targeting is an item-person map. Exhibit 37 presents an example of an item-person map of an item calibration. The crosses represent students and the numbers are items. In case of a partial credit item, each threshold is indicated after the item number. The vertical line represents the measured latent scale with high-performing students and difficult items at the top and low-performing students and easy items at the bottom. The response probability in this figure is 0.5, which means that students with an ability equal to the difficulty (or threshold) of an item have a 50% chance to respond correctly to that item. The figure shows that the test in this example is well-targeted, with the difficulties of the items a good match to the abilities of the students.
12.4.2.3 Assessment of item fit

A decision to review a test item should be based on a range of different criteria. Generally, an item would be flagged for review if the preliminary item calibration shows an infit statistic considerably higher than 1 (e.g., infit > 1.2) as well as low discrimination (e.g., item-rest correlation 0.2 or lower). Both item fit criteria as well as the content of the item should also be looked at when deciding to remove or retain a flagged item for scaling. See Chapter 4 for technical standards on item fit and discrimination.

In addition to assessing the item fit, the functioning of the partial credit score guides should be analysed further by reviewing the proportion of responses in each response category and the correct ordering of mean abilities of students across response categories.

12.4.2.4 Differential Item Functioning (DIF)

The quality of the items can be explored by assessing differential item functioning (DIF) by groups such as year levels, assessment cycles and gender. DIF occurs when groups of students with the same ability have different probabilities of responding correctly to an item. For example, if boys have a higher probability of success than girls with the same ability on an item, the item shows DIF in favour of boys. This constitutes a violation of the model, which assumes that the probability is only a function of ability and not of any other variable. Substantial item DIF with respect to gender may then result in bias of performance estimates across gender groups.

12.4.2.5 Assessment of anchor/link items

For the anchor test design, even if common items are carefully selected, those which do not behave well in the actual tests should not be used as the final anchor items to link the two tests. For example, in a test given to two groups of students at different year levels, one would expect the students at the higher year level would find a common item relatively easier than students at the lower year level. If the item was found to be harder for the students at the higher year level, that particular common item would be unsuitable for equating. Therefore, before equating, in addition to assessing individual item fit, a careful check should be carried out to assess the degree to which link items are working or behaving well across the two different tests. Below are some common checks to assess the link items.

- **Item-rest correlation**: The item-rest correlation between the two groups of students should be relatively similar. This check is done by computing the correlation between each item and the rest of the items across different groups. The item-rest correlation relates to the item’s discriminating ability in Classical Test Theory. Thus, this check aims to ensure that the item discrimination is stable across groups.

- **Item difficulty scatter plot** [14]: The scatter plot below compares paired item parameter estimates of common items between the two groups. It is used to evaluate the invariance of the relative item difficulty and the quality of items. Each item has an item parameter estimate and a standard error from each of two independent test calibrations. Using the standard errors, a 95% confidence interval can be constructed in the plot to investigate how satisfactorily the item points in the plot are positioned within the 95% confidence interval band.
If the same item behaves differently between two groups, the possible reasons of a change in an item must be investigated. These might include, for example, time difference between tests or different student samples taking the test.
12.5 Estimating student ability

Once the item difficulties are located on the Rasch scale, student scores can be computed. Education assessments can have two major purposes [15]:

- To measure the knowledge and skills of particular students. The performance of each student usually will have an impact on his or her future (e.g., school career, admission to post-secondary education, and so on). It is therefore particularly important to minimise the measurement error associated with each individual's estimate.
- To assess the knowledge or skills of a population. When the performance of individuals will have no impact on their school career or professional life, the goal of reducing error in making inferences about the target population is more important than the goal of reducing errors at the individual level.

This section discusses different types of student ability estimates, which can be categorised into two main groups: point estimates and plausible values. Exhibit 41 is the summary table of the different types of student ability estimates.

12.5.1 Point estimates

Maximum likelihood estimation (ML) [14], produces a single ability estimate based on the score obtained on a particular set of items. In other words, given the score on the test, a maximum likelihood estimate (MLE) of ability is made. For one-parameter logistic models (Rasch models), this method only requires the total score on the test rather than the actual response pattern to compute an ability estimate. (For two- or three-parameter logistic models, point estimates require the actual response patterns.) Another type of estimate, called the weighted maximum likelihood estimate (WLE) [16, 17], corrects some bias in the MLE procedure and also provides one ability estimate for each total score on the test. Both are discrete point estimates of a person parameter as shown in Exhibit 39, meaning that any raw score for a particular test will always be associated with one and only one MLE or WLE value.

Both MLE and WLE estimates are derived only from the item response model; they do not consider the population parameters and their variances, and require arbitrary treatment of perfect and zero scores. WLE is commonly used for contextual indices derived from the student or school questionnaire because it is a better estimate than MLE at the individual student level. WLEs will be unbiased only if the test is well-targeted. If the test is too easy, then the mean of the WLEs will be underestimated (the ceiling effect). If it is too difficult, the mean of the WLEs will then be overestimated (the floor effect).

Exhibit 39: Maximum likelihood estimates

---

28 Maximum likelihood estimation (ML) is a generic method. Different methods exist for maximising specific likelihood functions such as marginal likelihood (MML), conditional likelihood (CML), and joint likelihood (JML).
Population estimates are only unbiased if they are drawn from a model that includes all the information of interest. Therefore, it is very important to construct the model to take account of variables such as student and school backgrounds when drawing population estimates. The available student-level and school-level information, other than the student responses to the cognitive test, should be incorporated into the model. If there are subgroups within the population that make the assumption of normality no longer valid, the characteristics of these subgroups need to be taken into account [18]. This process is known as ‘conditioning’, in which variables that refer to the group (e.g., gender or language background) are included in the model. It is then possible to make correct inferences about the differences between subgroups based on these variables.

Marginal maximum likelihood (MML) estimation uses the student’s response patterns combined with other information about the population that the student is a member of to reflect the uncertainty in the estimate. When a student obtains a score on a test, that score is only an estimate of his or her ability. It can be assumed that the student’s latent ability is an effect which is sampled from a population with a well-defined distribution (e.g., normal). Hence, when taking account of marginal parameters\(^\text{29}\), the estimated distribution of student ability, known as the posterior distribution in Exhibit 40, is the most likely range in which a student’s true ability lies. This posterior distribution takes into account the item response pattern and the population model [18]. Note that the distribution is narrower in the middle of the scale and wider at the extreme ends of the scale, reflecting the greater uncertainty in ability estimates approaching perfect and zero scores.

**Exhibit 40:**
Marginal maximum likelihood estimates

The expected a-posteriori (EAP) estimate is the mean of the posterior distribution for each student. The posterior distribution is the same for all students with the same total score. Thus, the EAP estimation process provides the same estimate for students with the same total score. EAP estimates are therefore discrete. Although the mean of EAPs is an unbiased estimate of the population mean, the individual EAP values are biased towards the mean of the group and the variance of the EAPs is underestimated [19].

\[^29\] The likelihood that these marginal (or peripheral, with respect to student ability) parameters generate the observed data are then maximised, hence the term MML.
12.5.2 Plausible values

Plausible values (PVs) are random draws from the marginal posterior distribution of scores that could be reasonably assigned to each individual. The resulting distribution will be a better representation of the underlying continuous population distribution. It should be noted that they are not test scores and should not be treated as such. PVs contain random error variance components and are not optimal as scores for individuals [18]. In contrast to MLEs, WLEs and EAPs, the variance computed using PVs is not biased and therefore the percentiles based on PVs are also unbiased [18]. PVs perform better than WLE, MLE and EAP estimates as a description of the performance of the population [20].

A set of PVs, usually five, are drawn for each student for each scale or subscale. Population statistics should be estimated using each PV separately. The reported population statistic would be the average of each PV statistic. For instance, if analysis aims to consider the correlation coefficient between the social index and the reading performance, then five correlation coefficients should be computed and then averaged. Plausible values should never be averaged at the student level, such as by computing in the dataset the mean of the five PVs at the student level and then computing the statistic of interest using that average PV value.

Using one plausible value or five plausible values does not make a substantial difference in large samples. Because its distribution is unbiased, one PV would still perform better than using EAPs for population estimation. However, the computation of the standard error on one PV would only cover sampling error. To ensure the standard error accounts for both sampling and measurement errors, the computation must use five PVs. Therefore, it is recommended that during the exploratory phase of the data, statistical analyses can be based on a single PV. The reported results however should be based on five PVs, even on large samples. Results based on five PVs not only will guarantee consistencies between results but also are indisputable from a theoretical point of view [15].
### Exhibit 41: Summary of individual and population estimates

<table>
<thead>
<tr>
<th>Type of Estimates</th>
<th>Characteristics</th>
</tr>
</thead>
</table>
| Maximum likelihood estimate (MLE)                     | • One ability estimate for each total score on the test  
• Mean is unbiased if the test is well-targeted  
• Overestimates variance                                |
| Weighted maximum likelihood estimate (WLE)             | • One ability estimate for each total score on the test while correcting some bias  
• Better estimate than MLE at the individual student level  
• Mean is unbiased if the test is well-targeted  
• Still overestimates variance (but slightly better than MLE) |
| Expected a-posteriori estimate (EAP)                   | • Mean of the posterior distribution for each student  
• Mean is unbiased  
• Underestimates variance  
• Individual level estimates are biased towards the group mean  
• Takes into account the population (or sub-population) model |
| One plausible value (PV)                               | • A single random draw from the marginal posterior distribution of scores  
• Mean is unbiased  
• Variance is unbiased  
• Performs better than using EAPs for population estimate  
• Standard error on one PV would only cover sampling error  
• Recommended only for the exploratory phase of the data  
• Takes into account the population (or sub-population) model |
| A set of five plausible values                         | • Random draws from the marginal posterior distribution of scores  
• Better representation of the underlying continuous population distribution  
• Not optimal as scores for individuals  
• Mean is unbiased  
• Variance is unbiased  
• Standard error includes both sampling and measurement errors  
• Performs well as a description of the performance of the population  
• Recommended for the reported results  
• Takes into account the population (or sub-population) model |
12.5.3 Standard errors

The standard error is defined as the difference between the measured value and the true value. This difference is a component of the 'other variation' that is visualised in Exhibit 35. The sources of standard error can include noise or uncertainty components such as bad items, human error and mental states like student fatigue. It is therefore very important to estimate the standard errors to assess the precision of the parameter estimates. Standard errors in general analysis and various methods for calculating the standard errors of different statistics are discussed in more detail in Chapter 13.

It is important to note that because variances are biased for some estimators, residual variances will also be biased. For example, if WLEs estimates are used in a regression analysis, the standard error on the regression coefficients will also be biased.

For calculating standard errors of statistics based on PVs, special software is required. There are now complex survey analysis routines available in SAS and SPSS that can correctly calculate standard errors. In addition, ACER has created macros for SPSS and SAS, which can use replicate weights for calculation of standard errors as well.
12.6 Scale development

One of the most powerful ways to explain student achievement is through a scale which describes comparative abilities. Such a described scale provides details of the typical skills of the student with associated items at different points along the ability scale. The following section provides an overview of linear transformation and developing proficiency levels for a described scale.

12.6.1 Linear transformation

Item Response Theory software packages place items on a continuous scale measured in logits. After the equating and scaling processes, the scores in logits are transformed to a scale with a chosen mean and standard deviation by applying a linear transformation. This allows scores to be reported from a test on a readily understandable scale. For example, the India National Achievement Surveys are reported on a scale that originally had a mean of 250 and a standard deviation of 50; which is more readily interpreted than a mean of 0 (zero) and a standard deviation of 1 (where about half of the population would have negative scores). In conjunction with the use of standard errors, the use of a reporting scale is particularly helpful when providing information related to content, norm or reference groups and trends.

12.6.2 Generation of proficiency levels

One of the key objectives of assessment is to monitor trends in student performance over time. The assessment scale forms the basis for the empirical comparison of student performance. In addition to the metric established for the scale, a set of proficiency levels with substantive descriptions should also be established. These descriptive levels are syntheses of the item contents within each level. Comparison of student achievement against the proficiency levels provides an empirically and substantively convenient way of describing profiles of student achievement. Students whose results are located within a particular level of proficiency are typically able to demonstrate the understandings and skills associated with that level, and also typically possess the understandings and skills defined as applying at lower proficiency levels [21].

Establishing proficiency levels can be based on an approach used in many large-scale surveys such as the Programme for International Student Assessment (PISA) [22], Monitoring Trends in Educational Growth (MTEG) [23], the Trends in International Mathematics and Science Study (TIMSS) [24] and the National Assessment Program – Information and Communication Technology Literacy [19]. At the heart of this method is the notion that proficiency level can be interpreted consistently. A hypothetical example of a described proficiency scale for mathematics is shown in Exhibit 42.
Once the position of the proficiency levels on the scale is established, the process for setting standards can begin. The ‘proficient standard’ represents a point on the proficiency scale that represents a ‘challenging but reasonable’ expectation for students to have reached by the end of a relative year of study. The proficient standard can be set by going through a process using empirical judgemental technique. It requires stakeholders to examine the test items and the results from the assessments and agree on a proficient standard.
12.7 Summary

This chapter presents scaling methodology by introducing the concept from the perspectives of Classical Test Theory (CTT) and Item Response Theory (IRT). It discusses the rationale of choosing IRT over CTT for scaling and analysis methods. Different IRT models (one-, two- and three-parameter logistic models) for scaling are presented, discussed and compared.

The concepts of linking and calibration are introduced, and different options for these statistical processes are discussed. The calibration process leads to parameter estimation and estimating student ability. Various types of ability estimates are subsequently discussed. WLEs would be suitable at individual student level. For population estimates, using one plausible value is recommended for the exploratory phase of the data and a set of five PVs for the reported results.

Finally, this chapter wraps up with a discussion on the development of a common scale. This important process of scale development translates the test results into understandable and meaningful information for reporting.
Checklist: Scaling methodology

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Scaling model and linking design selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Has an appropriate Item Response Theory (IRT) model been selected for item analysis?</td>
</tr>
<tr>
<td>2. Has an appropriate linking design been selected?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Has the data been cleaned?</td>
</tr>
<tr>
<td>4. Has the data been re-coded where applicable?</td>
</tr>
<tr>
<td>5. Have the 'not-reached' items been identified and omitted from the calibration?</td>
</tr>
<tr>
<td>6. Have appropriate procedures been applied to ensure each sampled group contributes to the item statistics equally?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item review</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Has the model fit statistics of each item been assessed?</td>
</tr>
<tr>
<td>8. Is the test targeting the ability levels of the students well?</td>
</tr>
<tr>
<td>9. Have items been examined for evidence of differential item functioning (DIF)?</td>
</tr>
<tr>
<td>10. Have items that have poor psychometric characteristics been deleted or flagged?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test equating</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Have the anchor/common items been checked to see if they have stable characteristics across the different tests?</td>
</tr>
<tr>
<td>12. Has the set of anchor/common items been finalised?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Have the student ability estimates been drawn?</td>
</tr>
<tr>
<td>14. Is the choice of the type of ability estimate appropriate for the purposes of the assessment program?</td>
</tr>
<tr>
<td>15. Has the conditioning model for drawing plausible values (PVs) taken into account variables such as student and school background information?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale development</th>
</tr>
</thead>
<tbody>
<tr>
<td>16. Has the linear transformation been applied to transform the logit scores to a scale with a chosen mean and standard deviation?</td>
</tr>
<tr>
<td>17. Has a set of proficiency levels been developed to describe the established scale?</td>
</tr>
</tbody>
</table>
### Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor items</td>
<td>In measurement, anchor items are items which are selected to link two or more tests together. Once selected and having their item parameters estimated, these parameters are then taken as fixed during the linking process.</td>
</tr>
<tr>
<td>Latent trait</td>
<td>A trait that is not directly observable, such as maths ability. Also called latent construct, it needs to be derived from a set of observed or indicator variables (see the definition of 'indicator variable' in Chapter 7)</td>
</tr>
<tr>
<td>Logit</td>
<td>Log odd units. This unit is based on the logarithm of odds ratio of an event. The odds ratio is the probability for an event divided by the probability against an event. Logits have a mean of 0 and standard deviation of 1.</td>
</tr>
<tr>
<td>Parameter</td>
<td>A characteristic that defines a population, such as its variability or its average. A characteristic that defines a sample is called a statistic.</td>
</tr>
<tr>
<td>Point estimates</td>
<td>Estimates of parameters that relate to a single value of the corresponding statistic, often referred to as the ‘best guess of a parameter’. In the context of this chapter, point estimates refer to values of statistics that refer to a one member of a population, for example a student score.</td>
</tr>
<tr>
<td>Test targeting</td>
<td>In the context of test design, test targeting refers to the process in which item difficulties are matched with the ability levels of the target population.</td>
</tr>
</tbody>
</table>
References


Analysis for reporting
Analysis for reporting

This chapter discusses the types of analyses that can be conducted so that assessment results can inform policy and practice.

In Chapter 12, scaling using an Item Response Theory (IRT) framework was introduced as the primary data analysis method for large-scale assessments.

However, the results of a primary data analysis, while a necessary precursor, are not sufficient for reporting purposes. Large-scale assessments are often focused on the effects of other factors within the assessment results (e.g., whether class size affects student achievement). Analyses of the relationships among substantive variables of interest are often necessary beyond the primary scaling and data analysis. More often than not, policy level reports focus on these relational analyses. (See Chapter 14 for more details on reporting and dissemination.) This chapter provides overview of analytical methodologies that focus on exploring substantive relationships in the data for the purpose of reporting.
13.1 General methodology

Before conducting any analysis for reporting, it is essential to select the suitable analysis software and prepare a data analysis plan.

13.1.1 Analysis software

Computer-based statistical packages are important tools for researchers in education. These software packages have reduced the time and effort needed for researchers to be able to undertake many of the complex calculations that are required for specialised statistical analyses. It should be emphasised, however, that these packages have not reduced the need for researchers to understand the assumptions behind statistical analyses, the potential applications and limitations of various statistical tests, and methods for interpretation of their results. There are many statistical analysis software packages available in the market. The most widely used commercial packages are STATA, SAS and SPSS. All three statistical software packages are proficient and flexible in analysing complex survey data. Many open source and free options are also available, ranging from basic (Statistics Open For All - SOFA) to comprehensive (eg R Package).

13.1.2 Data analysis plan

The data analysis plan contains a description of the research questions and what the various steps in the analysis are going to be. The plan acts as a starting point for the analysis. An analysis plan should be established as early as possible, ideally at the test design phase. Developing an analysis plan will not only help to inform the design but also assist in defining the data that need to be collected. After the data collection stage, the analysis plan should be refined. It ensures that the analysis can be carried out in a targeted manner. Before establishing an analysis plan, it is important to consider the questions as outlined in Exhibit 43 [2].

Exhibit 43:
Questions for consideration when establishing an analysis plan

<table>
<thead>
<tr>
<th>Questions</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the practical purpose of the analysis?</td>
<td>To: find solutions, build theory, develop an intervention, evaluate something, inform subsequent data collection, and inform ongoing data collection.</td>
</tr>
<tr>
<td>What is the analytic purpose?</td>
<td>To: identify, explore, explain, compare, differentiate, confirm, or some combination of these purposes.</td>
</tr>
<tr>
<td>How is the analysis connected to the research question(s)?</td>
<td>The analysis should directly inform one or more of the research questions.</td>
</tr>
<tr>
<td>What is the program timeline?</td>
<td>Consider how quickly the findings from the analysis are needed. The answer can range from ‘tomorrow’ to ‘no foreseeable deadline’.</td>
</tr>
<tr>
<td>What resources are available?</td>
<td>For example, consider the number and skills of analysts available to work on the project and the analysis software that is available.</td>
</tr>
</tbody>
</table>
Once there is a good sense of the analysis purposes, the quality of the data, and the resources and time available, an analysis plan can be developed. Below are some items that need to be considered in developing an analysis plan [2].

Specify how many separate analyses will be conducted and the timeline for each. For each separate analysis specify:

- which research question(s) it will inform and how
- precisely which data will be used, including a list of inclusions, exclusions, and missing criteria if required
- how many people will be involved in the analysis and their specific roles
- the primary analytic purpose, such as to identify, explore, explain, confirm, compare (note: the verb used is very important, so choose carefully)
- how codes will be created and defined (see Chapter 11 for more details on developing a codebook)
- rules for applying codes to the data (eg will all text be coded?)
- how coding reliability will be established, including reconciling discrepancies
- which data reduction techniques, if any, will be applied
- which between-group comparisons, if any, will be made, and how this will be done
- how data from different data collection methods will be integrated
- what is expected as a research output (eg in-house report, chapter in public report, table shells).

<table>
<thead>
<tr>
<th>Questions</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>How large is the dataset?</td>
<td>The size of a qualitative dataset can range from a few in-depth interviews or focus groups to hundreds of various types of data collection activities (such as observation interviews, focus groups and secondary data).</td>
</tr>
<tr>
<td>How heterogeneous are the data types?</td>
<td>The data may come from only one type of activity, such as in-depth interviews, or in a larger study, data may be from focus groups, observations and secondary sources as well. There may also be quantitative data relevant to the analytic objectives.</td>
</tr>
<tr>
<td>Which data should be used for a particular analysis?</td>
<td>The answer to the question can range from a very small section of text to an entire dataset, and depends on the overall size of the dataset, research objectives and time constraints.</td>
</tr>
<tr>
<td>Who is the audience for the analysis, and how will members judge the process and subsequent findings?</td>
<td>There may be one primary audience for the report, or several different audiences. Think carefully about for whom a particular analysis is intended.</td>
</tr>
</tbody>
</table>


13.2 Relationship between achievement and background variables

Contextual student variables that are associated with educational outcomes are important to consider when examining student achievement in the context of large-scale assessments [3, 4, 5]. To increase educators’ understanding of student and school performance and to support efforts to increase such performance, it is important to include, and to better understand, the associations of such factors with measures of student achievement.

One major interest in large-scale assessments is to examine the extent to which student performance relates to other variables. These variables could include:

- attitude indices (e.g., interest and enjoyment in mathematics)
- questionnaire scale indices which are constructed by several items (e.g., socio-economic background)
- student-level variables (e.g., gender, language spoken at home)
- school-level variables (e.g., school management, urban/rural location).

Attitude and questionnaire scale indices are often constructed with several items. For example, social-economic background index could be made up by combining parents’ occupation, parents’ education and an indication of possessions in the home indicating wealth.

With appropriate analysis, the results can provide useful information to inform policy questions and recommendations. The results can also provide insights into the relationships between educational policies and practice. For example, the results could be used as a value-added estimation or an evaluation tool for an educational intervention.

In large-scale educational surveys, questionnaires are often used to collect information from schools, teachers, parents and students. The data collected are either used to construct different questionnaire scale indices or as aggregated indicators. The indices and indicators can then be used to conduct analysis at the student level. Below are a few examples of how the information could be utilised:

- mean score in student performance, by parental education
- percentage of students at each proficiency level, by language spoken at home
- performance on the mathematics scale, by quartiles of the mathematics self-efficacy scale
- growth in mathematics achievement over several assessment cycles, analysed in relation to school funding
- correlation between mathematics achievement and students’ socio-economic background.
13.3 Sample weighting

In most situations, samples do not precisely represent the population and therefore the population estimates derived from them would be biased (either an under- or overestimate) due to this misrepresentation. A correction technique, through the use of sampling weights (see Chapter 9), can be used to adjust the sample and reduce the bias in the population estimates.

Any analysis should, therefore, always be weighted at any stage of the process, whether it is the primary or secondary data analysis.

Weights are assigned to each student and to each school for the following reasons:

• students and schools in a particular state or region do not necessarily have the same probability of selection
• varying participation rates according to certain types of school or student characteristics require various non-response adjustments
• some explicit strata may be over-sampled for national reporting purposes.
13.4 Standard error and replication techniques

A standard error is the spread or variability of a sample statistic around its mean. In other words, it is a measure of the accuracy of a sample statistic as an estimate of an unknown population parameter. The variance of a statistic of interest, such as mean (\(\hat{\theta}\)), is computed as follows:

\[
\sigma^2_{(\hat{\theta})} = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\theta}_i - \hat{\theta})^2
\]

where \(n\) is the total sample size.

The standard error of the statistic \(\hat{\theta}\) is the standard deviation of the sampling distribution and is computed as the square root of the variance of the statistic of interest.

\[
\sigma_{(\hat{\theta})} = \sqrt{\sigma^2_{(\hat{\theta})}}
\]

The formulae above assume simple random sampling. For large-scale assessment, using a complex sample design is more cost- and time-effective than a simple random sample. However, students clustered within a school usually have more common characteristics than students from different schools. Therefore, the student samples cannot be considered as independent observations. It is not appropriate to use the above formulae directly to obtain standard errors of population estimates.

It is very important to report accurate and unbiased standard error, as these estimates are used for calculating the statistical significance of differences between subgroups. Standard errors for large-scale assessment are a combination of the sampling error and measurement error. To achieve unbiased standard errors from survey studies, the analysis must have accurate estimates of both sampling variance and measurement variance [8].

In the case of a complex sample survey, replication techniques should be applied when calculating the sampling variance of population estimates using replication methods. These methods enable a better estimate of the true sampling variance, which means that the standard errors are more accurate. Several replication methods are available to the analyst including jackknife repeated replication (JRR or JK1) and balanced repeated replication (BRR). A variant of JRR known as jackknife 2 (JK2) is often employed in large-scale surveys and recommended here because it is a simple and effective method. Computing the variance of a statistic of interest obtained through replication is as follows:

\[
\sigma^2_{(\hat{\theta})} = K \sum_{i=1}^{G} (\hat{\theta}_i - \hat{\theta})^2
\]

Where \(G\) is the number of replicates or primary sampling units and \(K\) is a constant based on the replication method used. For JK1, \(K = (G - 1) + G\); for JK2, \(K = 1\); for BRR, \(K = 1 + G\); for Fay’s variant, \(K = 1 + (G(1 - k)^2)\), where \(k\) is an inflating factor between 0 and 1 [6].
Unbiased measurement variance can be derived by using the plausible values method outlined in Chapter 12 and calculating the variance among the five plausible values. In addition, an equating error should be added as a third component of the standard error for comparison of assessment cycles.

In a report, each population estimate should be accompanied either by its confidence interval or standard error. In addition, tests of significance for the difference between estimates of subgroups should be provided, in order to describe the probability that differences are just a result of sampling and measurement error. Below are examples of significance tests for achievement mean differences in population estimates to be reported:

• between states and union territories
• between class levels
• between student background subgroups (eg social category)
• across assessment cycles.

Further discussion of calculating standard error of mean differences for subgroup analysis will be covered in the next subsections.

13.4.1 Mean differences between states and year levels

Pair-wise comparison charts allow the comparison of population estimates between one state and another, or between different year levels. Differences in means are considered significant when the test statistic $t$ is outside the critical values $\pm 1.96 \ (\alpha = 0.05)$. The $t$ value is calculated by dividing the difference in means by its standard error that is given by the formula:

$$ SE_{\text{diff ij}} = \sqrt{SE_i^2 + SE_j^2} $$

where $SE_{\text{diff ij}}$ is the standard error on the difference and $SE_i$ and $SE_j$ are the standard errors of the compared means $i$ and $j$. The standard error on a difference can only be computed this way if the comparison is between two independent samples like states or year levels. Samples are independent if they were drawn separately.

13.4.2 Mean differences between dependent subgroups

The formula for calculating the standard error provided above is only suitable when the subsamples being compared are independent [6]. In the case of dependent subgroups, the covariance between the two standard errors needs to be taken into account and the appropriate repeated replication technique (such as jackknife repeated replication) should be used to estimate the sampling error for mean differences. Because subgroups other than state and year level are dependent subsamples (eg gender, language background, parental highest education and school social-economic subgroups), the difference between statistics for subgroups of interest and the standard error of the difference can be derived by applying the replicate procedures using SPSS or SAS tailored macros. Differences between subgroups are considered significant when the test statistic $t$ is outside the critical values $\pm 1.96 \ (\alpha = 0.05)$. The $t$ value is calculated by dividing the mean difference by its standard error.
13.4.3 Mean differences between assessment cycles

Comparisons of assessment results across cycles can also be considered for reporting. As the equating process for the tests across the cycles introduces some additional errors into the calculation of any test statistic, an equating error term should be computed and added to the formula for the standard error of the difference, such as difference between cycle means.

When testing the difference of a statistic between the two assessments, the standard error of the difference is computed as follows:

\[
SE(\mu_{cycle\ B} - \mu_{cycle\ A}) = \sqrt{SE_{cycle\ B}^2 + SE_{cycle\ A}^2}
\]

where \( \mu \) can be any statistic in units on a cognitive scale (mean, percentile, gender difference, but not percentages) and \( SE \) is the respective standard error of this statistic\(^{30}\).

\(^{30}\)An additional source of error arises when cycles are linked with a subset of items. This is known as linking error. For discussion of this, refer to Monseur and Berezner [7] and OECD [8].
Regression analysis

Many educational research questions concern the relationship between student performance and other variables. These variables can be questionnaire indices and/or background variables. Regression analysis refers to a set of techniques for predicting an outcome variable using one or more explanatory variables. It involves creating a model for estimating or predicting one variable based on the values of others. This section will briefly discuss simple linear regression and multiple regression.

13.5.1 Simple linear regression

Linear regression is a probabilistic model that reflects the relationship between the dependent (or outcome) variable and the independent (or predictor) variable. It allows the researcher to use the pattern of previously collected data to build a predictive model that predicts the value of the outcome variable based on the value of the predictor variable \( x \). The most basic form of regression analysis is simple linear regression.

\[ Y = \alpha + \beta x + \epsilon \]

Where:

- \( Y \) is a dependent or outcome variable (e.g., student performance).
- \( x \) is an independent variable or predictor. It can be a continuous explanatory variable such as social economic index, or it can be a control or dummy variable (e.g., gender or school type).
- \( \alpha \) is the \( Y \)-intercept parameter. It is the theoretical estimate of the outcome variable when \( Y \) is zero.
- \( \beta \) is the slope parameter (typically called the regression coefficient). It is a measure of linear association between variables \( Y \) and \( x \), indicating the strength and direction between two variables.
- \( \epsilon \) is the random variable representing an error or residual of the actual (observed) values of \( Y \).

Simple linear regression can be seen as the process of drawing a line to represent an association between two variables on a scatter plot. For example, Exhibit 44 shows the scatter plot of values \( x \) and \( y \). The black line, which is called the regression (or fit) line, is drawn based on the regression model using the observed data of the scatter plot. Under this regression model, for one unit of \( x \), \( y \) changes to an average of 0.6565 (regression coefficient). \( R^2 \) is a statistical measure of the amount of variance explained by the model, and indicates how close the data are to the fitted regression line. In other words, it is the percentage of the response variable variation that is explained by a linear model. Here, 57.29% of the variation is explained by the model.
A study was done for the Programme for International Student Assessment (PISA) 2006 on how socio-economic background related to student performance in science [9]. In the regression model in this study, mean scores in science are the outcome variables, and the explanatory or predictor variable is the PISA index of economic, social and cultural status (ESCS).

Exhibit 45 shows the results of France, Germany and The United Kingdom. The last column on the left of the table corresponds to the model’s regression coefficients for each country. It shows how much student performance in Science changes with a change of one unit on the PISA index of ESCS. France has the steepest slopes, where one unit of increase on the PISA index of ESCS is associated with an average increase of 54 score points on the science performance scale. On the other hand, Germany and the United Kingdom have a similar slope, with one unit of PISA index of ESCS to an average of 46 and 48 score points on the science performance scale respectively.

However, in the United Kingdom, there are many students from disadvantaged socio-economic backgrounds still achieving well, while many students from advantaged background performed lower than predicted. Exhibit 45 also shows the percentage of variance in student performance explained by the model – this is equivalent to $R^2$. The modelled relationship, therefore, only explains 13.9% of the performance variation in the United Kingdom. As can be seen in the exhibit, in Germany, 19.0% of the performance variation is explained by socio-economic background, indicating student performance in Germany was predicted by ESCS more closely.
Below are some further examples of research questions to which simple linear regression can be applied:

- Does student self-efficacy in mathematics influence the likelihood of obtaining a higher mathematics score?
- Are some school types more effective than others?
- Does a parent’s highest education level have an impact on a student’s educational progress?

13.5.2 Multiple regression

Often models with more than one independent variable are needed to deal with complex phenomena. Linear regression can then be developed further to represent an association between a continuous scale and more than one continuous explanatory or control variable. Multiple regression is a statistical procedure to analyse the relationship between a dependent variable and two or more independent/predictor variables. The most important uses of the technique are:

- to find the best linear prediction equation and evaluate its accuracy
- to control for other confounding factors in order to evaluate the contribution of a specific variable or set of variables
- to find structural relations
- to estimate population parameters and test hypotheses.

In making inferences about linear combinations of the model parameters $\beta_1, \beta_2, \ldots, \beta_k$, the dependent variable $Y$ is modelled as:

$$ Y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon $$

There is no limit on the number of independent variables that can be included in a regression analysis. However, there must always be many more cases than there are independent variables. Further, the more independent variables that are included in the model, the more difficult it can become to meaningfully interpret the model; in fact, a model with a large number of disparate independent variables is essentially meaningless, for practical purposes.

The multiple regression coefficient, $\beta_k$, of an independent variable indicates the expected change of the dependent variable, $Y$, for a unit increase of the independent variable, holding constant or ‘controlling’ for all other variables in the regression model.
13.5.3 Hierarchical regression

Hierarchical multiple regression\(^3\) is used to examine the unique effect of an independent variable on a dependent variable after allowing for the effect of other independent variables. Hierarchical regression is done by analysing several regression models in a hierarchical manner, adding a set of predictors in each successive models. The structure of the hierarchy is dependent on the study design. The set of predictors that needs to be partialled out (or controlled for) is added first. The change in \(R^2\) from one hierarchical model to the next represents the amount of unique variance explained above and beyond that explained by the previous models.

This is a particularly useful technique in large-scale surveys where there are a number of variables or conditions which at first glance seem to be associated with educational outcomes but where in fact the association disappears after partialling out other, often more fundamental, variables.

For example, in the Afghanistan Monitoring Trends in Educational Growth (MTEG) 2013, regression analyses were conducted to investigate the relationship between achievement and various background factors such as degree of rurality. At first there appeared to be a positive association between being schooled in a more urban environment and educational achievement – students in urban environments achieved better than students in rural environments. However when socio-economic status (SES) was included in the multiple regression model, the partial or unique effect of being schooled in a more urban environment disappeared – students from higher SES backgrounds do better than those from lower SES backgrounds no matter where they are schooled; and students from low-SES backgrounds schooled in urban environments achieve the same as students from low-SES backgrounds schooled in rural environments \(^{10}\).

\(^3\) Also called step-wise regression, and not to be confused with hierarchical linear models (HLM). For hierarchical linear models, see Section 13.6.
13.6 Multilevel analysis

Linear regression models may provide an incomplete or misleading representation of efficiency in education systems. This can be because linear regression models do not take into account the potential effects of clustering in a complex sample design in an education survey.

Multilevel analysis uses the hierarchical linear model [11]. The hierarchical linear model is a type of regression analysis for multilevel data (e.g., students within schools, schools within districts, districts within states). Multilevel data is a random coefficient regression analysis for data with several nested levels, where the dependent variable is at the lowest level. Multilevel analysis is a suitable approach to take into account the social contexts as well as the individual respondents or subjects [12]. It can be used on data with many levels, although two-level models are the most common. For models with more levels, it can, however, become more difficult to interpret the results.

For example, in order to help policymakers to know where to target a reform, a multilevel analysis was done in PISA 2009 Plus [13], to investigate the relationship between social segregation (i.e., between-school variance in students’ socio-economic background) and academic segregations (i.e., between-school variance in performance). Because these are closely related, multilevel analysis becomes useful to examine whether social segregation is an outcome of academic segregation or vice versa. If, for example, a significant proportion of students are enrolled in schools with high admission fees, academic segregation could be partly an outcome of social segregation. But if students were assigned to schools based on their learning performance, social segregation may be an outcome of academic segregation. Based on the analytical results, policymakers would then be able to tailor an educational reform to serve the specific needs to the education system.

The PISA index of economic, social and cultural status (ESCS) is used as an indicator of student socio-economic background. To estimate the importance of the school socio-economic composition effect on student performance, it also requires the school average of the PISA student ESCS (MU_ESCS). The following multilevel model is used to estimate the composite effect.

\[ Y_{ij} = \beta_{0j} + \beta_{1j} (ESCS) + \varepsilon_{ij} \]
\[ \beta_{0j} = \gamma_{00} + \gamma_{01} (MU_{ESCS}) + U_{0j} \]
\[ \beta_{1j} = \gamma_{10} \]

Taking state Himachal Pradesh as an example, the between-school variances of the empty model and the model with individual and school-level ESCS are 1741 and 1285 respectively. Therefore, the student and school socio-economic backgrounds together explain about 26% of the school variance, that is, \((1741 - 1285) / 1741\). This suggests that there is less difference between schools once ESCS is accounted for, but also indicates there are other important factors influencing differences between schools as there is considerable between school variance (74%) that is not accounted for by ESCS. Further discussion on multilevel analysis in PISA Plus 2009 is in the PISA 2009 Plus results [13].
13.7 Summary

This chapter discusses the types of analyses that focus on the relationships across the variables, to inform policy and practice with respect to the research goals of an assessment program. The importance of having a data analysis plan prepared as early as possible, preferably at the design stage, is emphasised. It also discusses key issues that need to be considered when preparing the analysis plan.

The chapter also provides an overview of sample weighting and the computation of standard errors on differences. It briefly discusses why the covariance should be taken into account for comparisons of dependent subgroups and for comparisons between assessment cycles. The methodological focus of the chapter is on regression analysis – including simple linear, multiple regression, and multilevel analysis methods – as the regression analysis approach is common in educational research.
### Checklist: Analysis for reporting

If you are planning or implementing a large-scale assessment, have these points been addressed?

<table>
<thead>
<tr>
<th>Analysis software</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Has appropriate statistical analysis software been selected and obtained?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data analysis plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Has a data analysis plan been developed (preferably during the test design phase)?</td>
</tr>
<tr>
<td>3. Has the data analysis plan been refined after the data collection phase?</td>
</tr>
<tr>
<td>4. Does the final data analysis plan take into account considerations such as the analysis purpose, time, resources, timeline, data type and audience?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Have all variables and indices been prepared for reporting?</td>
</tr>
<tr>
<td>6. Have the final student weights and replicate weights that are appropriate for the study design been finalised and applied for analysis?</td>
</tr>
<tr>
<td>7. Have the standard errors of the statistics been calculated according to the nature of the analysis and sample design?</td>
</tr>
<tr>
<td>8. Has any linear regression analysis been considered?</td>
</tr>
<tr>
<td>9. Has any multiple regression analysis been considered?</td>
</tr>
<tr>
<td>10. Has any multilevel analysis been considered?</td>
</tr>
<tr>
<td>11. Have any additional analyses been considered?</td>
</tr>
</tbody>
</table>
### Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data reduction</td>
<td>Methods to convert the data into a more compact form. For example, converting texts (words or sentences) into numeric codes.</td>
</tr>
<tr>
<td>Index</td>
<td>In the context of this chapter, an index is a scaled indicator of a measure that is composed of several values or other measures. For example, a socioeconomic index might be composed of income, health factors, education level and other components.</td>
</tr>
<tr>
<td>Outcome</td>
<td>In regression analysis, the outcome variable is the variable of interest which changes (or varies) depending on the predictor variables.</td>
</tr>
<tr>
<td>Predictor</td>
<td>In regression analysis, predictors are variables that are used in the regression model as explaining the outcome variable.</td>
</tr>
<tr>
<td>Variance explained</td>
<td>The extent to which a regression model (or any model in general) explains the variation in the outcome variable. The amount of variance explained is proportional to how well the model predicts or explains the data.</td>
</tr>
</tbody>
</table>
References


CHAPTER 14

Reporting and dissemination
Reporting and dissemination

This chapter will briefly discuss some of the important issues to consider for the reporting and dissemination of assessment information.

This chapter begins by discussing the importance of communicating the assessment results effectively to stakeholders to enable them to understand and use the results. It then discusses the need to develop a dissemination strategy which involves considering what dissemination methods will be most effective. A range of dissemination methods are outlined, before more in-depth information is provided about an essential dissemination approach: report writing. Some of the factors which influence the likelihood of the assessment results having an impact on policies are outlined, and lastly, some of the potential policy impacts are discussed. More in-depth information about policy goals and issues can be found in Chapter 2.

While this chapter appears at the end of the handbook, this is not to imply that decisions regarding dissemination should be left until the end of an assessment program. Dissemination activities, including reporting, should be planned for in program budgets and timelines. Consideration of what information will be of value to stakeholders needs to occur alongside the assessment and data analysis design, to ensure that it will be possible to report on this information. If assessment programs are to have an impact on policy and practice, consultation with stakeholders is most effective if it occurs throughout the assessment program, rather than just at the conclusion of the assessment.

14.1 Understanding the results

As is discussed in Chapter 2, national and state learning assessments can provide valuable information on student outcomes and should be designed to inform educational policy and practice. The data collected from learning assessments can provide stakeholders with information on the education system with regards to issues of access, quality, efficiency and equity [1]. This information can then be used by stakeholders to make decisions about educational policies, with the ultimate aim of improving learning outcomes for students.

In order for stakeholders to make use of assessment results, it is essential that the information is reported on and disseminated appropriately. If the assessment is to increase stakeholder understanding, stakeholders need to have access to the assessment results, and the results should be presented in a way the stakeholders can comprehend. However, if learning assessments are to bring about change, stakeholders also need to understand the relevance of the results, be able to use the data to identify appropriate actions, and be in a position and system where they are able to enact change.

The reporting and dissemination approaches used by the assessment team, therefore, need to support stakeholders in understanding and making effective use of the assessment results. The next section discusses the need to develop a dissemination strategy that incorporates a range of methods to cater for diverse audiences.
14.2 Dissemination strategy

A dissemination strategy involves considering the intended audiences, the different interests and requirements of those audiences, and the dissemination approaches (including reporting options) that will meet the target audiences’ needs.

14.2.1 Target audiences

The first step in developing a dissemination strategy is to identify the groups and individuals to whom the assessment results need to be communicated, in order for the assessment program to have the desired impacts. Generally it is preferable for the assessment results to be disseminated widely so that understanding is increased amongst a wide range of stakeholders at the national, state and district levels. There will also be key groups and individuals who are in positions to bring about change will be particularly important as target audiences.

When communicating assessment results, there is no one-size-fits-all approach, as stakeholders vary in aspects such as their areas of interest, technical understandings, time availability and political connections and influence. The assessment results need to be communicated in a way that meets the needs of the target audiences. For example, detailed information and descriptions about students’ proficiency levels would be useful information for curriculum planning organisations such as the National Council of Educational Research and Training and State Council of Educational Research and Training. Assessment information collected on teachers’ education and professional development, and the relationship between these factors and student achievement, would likely be of interest to teacher training organisations such as the National Council for Teacher Education. Information about the types of items that tended to be difficult for students could be of interest to teachers and school heads.

The best way to identify the different needs of stakeholders is through consultation. Consultation throughout the assessment program can be used to establish what the relevant policy and practice issues are, to gain a deeper understanding of stakeholders’ requirements, and also to develop stakeholders’ understanding and support of the assessment program. In India, stakeholders are likely to be geographically dispersed, so options such as teleconferences and webinars may need to be explored.

14.2.2 Dissemination methods

After the different stakeholders and their requirements have been identified, the next step is to think about the dissemination methods that are most suitable, taking into account the available resources. Reports are the most common way of communicating assessment results, and while they are an important dissemination method, they are unlikely to be read by all of the target audiences. Generally, a variety of activities are needed in order to reach stakeholders at all levels.
When thinking about dissemination approaches, consider the following questions for each method:

- Who are the target audiences?
- What are the benefits and potential drawbacks?
- How will the information reach stakeholders?
- What is the desired impact?
- What is the likelihood of the approach having an impact?
- How will the impact be measured?
- What are the resources available, and is the approach cost-effective?
- When should this dissemination approach be used? For example, is there an ideal time, or is the activity dependent on another activity being completed?

Exhibit 46 provides brief descriptions of a selection of dissemination activity options and some of their purposes. Reporting is discussed in more detail in the next section.

**Exhibit 46:**
Dissemination methods

<table>
<thead>
<tr>
<th>Dissemination method</th>
<th>Description</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main report</td>
<td>Report that covers all of the core aspects of the assessment. Almost all assessment projects have a main report.</td>
<td>Provides an overview of all aspects of the assessment so that a variety of stakeholders can understand the purpose, approach taken, results and implications.</td>
</tr>
<tr>
<td>Thematic reports</td>
<td>Reports that provide more detailed information than the main report around a particular topic of interest (e.g., a report on differences in achievement patterns between girls and boys in Delhi).</td>
<td>Provides additional detail that is of particular interest to a group of stakeholders. Producing thematic reports can help raise awareness about a particular priority area.</td>
</tr>
<tr>
<td>Summary reports, pamphlets</td>
<td>A summary of the important points from the main report. These reports can vary in length. An example might be a summary brochure on the learning achievement patterns of Indian students at Class V.</td>
<td>May be produced for a variety of stakeholders, including teachers, policymakers, the general public or key interest groups. Summary reports or pamphlets provide a fast way for stakeholders to learn about the important assessment results.</td>
</tr>
<tr>
<td>Technical report</td>
<td>Provides detailed information about the assessment processes and results. This information may be included as part of the main report or as a separate technical report.</td>
<td>Provides detailed information for stakeholders (such as researchers) to judge the quality of the assessment and to understand how the results should be interpreted. It also provides a record of activities which can inform future assessment phases. If a separate technical report is produced, this can mean some technical details can be left out of the main report making it more accessible for some stakeholders.</td>
</tr>
<tr>
<td>Assessment framework report</td>
<td>Provides details about the assessment framework that guided the development of the assessment (see Chapter 5). A summary of the assessment framework may be included in the main report; however, the full assessment framework may be published as a separate report.</td>
<td>Provides detailed information for stakeholders (such as researchers) to understand the purpose of the assessment and what is measures. It also provides a record of the procedures and definitions on which the assessment is based, which can inform future assessment phases.</td>
</tr>
<tr>
<td>Interim report</td>
<td>Provides information about how the program is proceeding and may provide some initial results prior to the main report being developed.</td>
<td>Provides information to stakeholders in a timely manner and can help stakeholders to feel engaged in the program.</td>
</tr>
<tr>
<td>Briefings</td>
<td>Short briefings providing a summary of the main information and possible implications, often for ministers and policymakers. Can be written or else delivered by presentation.</td>
<td>Main messages can be communicated concisely to decision-makers who do not have time to read a full report. These decision-makers can use this information to identify possible next steps.</td>
</tr>
</tbody>
</table>

---

32 For a more detailed discussion see Volume 5 of Using the Results of a National Assessment of Educational Achievement [1]. The five volumes of the World Bank’s series on National Assessments of Educational Achievement are available electronically from https://openknowledge.worldbank.org/handle/10986/2143
<table>
<thead>
<tr>
<th>Dissemination method</th>
<th>Description</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media reports</td>
<td>Can include newspaper articles, radio or television reports, blogs, videos and press conferences.</td>
<td>Allows information to be spread to a wider audience in an accessible way. However, care must be taken as the media may greatly simplify assessment results or focus only on more controversial results (e.g. league tables).</td>
</tr>
<tr>
<td>Press releases</td>
<td>Short written statements provided to the media.</td>
<td>Allows information to be spread to a wider audience in an accessible way. Allows for more control over what is reported by the media.</td>
</tr>
<tr>
<td>Assessment database</td>
<td>Assessment data that is publicly available or that is available to certain stakeholders/organisations that have been granted access.</td>
<td>Stakeholders, including government officials, researchers and organisations, can conduct secondary analyses using assessment data. Assessment data can be used by stakeholders to investigate particular areas of interest.</td>
</tr>
<tr>
<td>Conferences and workshops</td>
<td>Discussion and presentation of assessment to stakeholders. Workshops generally involve a smaller number of people and are more participatory than conferences.</td>
<td>Provides an opportunity to gather feedback from stakeholders and to discuss possible policy implications. May be held with particular interest groups, such as teacher trainers or curriculum developers, or may be for multiple stakeholders to bring together different perspectives and to reach agreement regarding next steps.</td>
</tr>
<tr>
<td>Websites or blogs</td>
<td>A webpage for the assessment program may contain links to different dissemination outputs. For example, reports, press releases and the assessment database may all be found on a webpage.</td>
<td>Allows stakeholders to easily find a range of assessment information in one place and to learn about project updates.</td>
</tr>
</tbody>
</table>

There are benefits and drawbacks to all dissemination activities, so anticipating likely issues or questions and preparing for these is essential. For example, an assessment team may develop a communication strategy, which may include nominating a media spokesperson and reaching agreement about what messages will be conveyed.

The use of simple language and clear messages in all dissemination activities will aid understanding, and technical information should be available to substantiate all statements made.
14.2.3 Dissemination plan

Decisions about the target audiences and the appropriate dissemination methods can be recorded in a product specification sheet and used to form a dissemination plan. The following information may be documented for each activity:

- product name (e.g., main report)
- summary description of the product
- intended users
- product priority
- detailed product description
- key production activities, responsibilities and time schedule
- production costs
- projected release date
- product price
- dependence on other products or inputs [2, p. 48]

A hypothetical example of a completed product specification sheet is shown in Exhibit 47.

**Exhibit 47:** Example product specification sheet

<table>
<thead>
<tr>
<th>Product name</th>
<th>National Achievement Survey (NAS) Class V Technical Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary description of the product</td>
<td>This report will provide technical information on all aspects of the Class V assessment program.</td>
</tr>
<tr>
<td>Intended users</td>
<td>This report will be used by evaluation/assessment experts. It will also provide documentation for internal and external team members involved in NAS to refer to in future assessment cycles. This includes those working on Class III, V, VIII and X NAS.</td>
</tr>
<tr>
<td>Product priority</td>
<td>High</td>
</tr>
<tr>
<td>Detailed product description</td>
<td>The technical report will include chapters on the following:</td>
</tr>
<tr>
<td></td>
<td>• item development and test design</td>
</tr>
<tr>
<td></td>
<td>• questionnaire development</td>
</tr>
<tr>
<td></td>
<td>• instrument translation</td>
</tr>
<tr>
<td></td>
<td>• sampling and weighting</td>
</tr>
<tr>
<td></td>
<td>• field operations</td>
</tr>
<tr>
<td></td>
<td>• data processing</td>
</tr>
<tr>
<td></td>
<td>• scaling and analysis of the assessment data and the contextual data</td>
</tr>
<tr>
<td></td>
<td>• assessment findings, limitations and recommendations.</td>
</tr>
<tr>
<td>Key production activities, responsibilities and time schedule</td>
<td>August/September 2015 – Drafts of each chapter written by each lead authors</td>
</tr>
<tr>
<td></td>
<td>October 2015 – Report editors to review chapters</td>
</tr>
<tr>
<td></td>
<td>October – November 2015 – Authors revise chapters</td>
</tr>
<tr>
<td></td>
<td>December 2015 – Report formatted and proofread</td>
</tr>
<tr>
<td>Production costs</td>
<td>To be confirmed</td>
</tr>
<tr>
<td>Projected release date</td>
<td>31st December 2015</td>
</tr>
<tr>
<td>Product price</td>
<td>No cost, freely available to download from website</td>
</tr>
<tr>
<td>Dependence on other products or inputs</td>
<td>The technical report will be produced after the main report, once all of the results have been analysed and reported on.</td>
</tr>
</tbody>
</table>

The information in the product specification sheets can be used to determine dissemination priorities, timelines and budgets. A dissemination plan should be developed early on in an assessment program, so that dissemination activities can be planned to occur throughout the project and maximise stakeholder engagement. Flexibility is also needed – the dissemination plan should be adapted as new stakeholders are identified, as resources become available, as evidence is gathered about what dissemination methods are most effective and as other possible policy implications of the assessment results become apparent.
For a population-based assessment, the dissemination plan will often include reporting individual school (and possibly individual student) results to school leaders. The assessment team may need to provide schools with support in interpreting the results. Consideration needs to be given as to whether the results are publicly reported at the school level. Reporting results at the school level can often result in the publication of league tables, leading to the creation of a high-stakes assessment. Assessments which are high stakes and which are tied to funding decisions can have unintended negative consequences such as resistance from schools to participate, a narrowing of the curriculum and teaching to the test [1].
14.3 Report writing

Printed and online reports may include a main report, summary reports or pamphlets, thematic reports and a technical report. The National Council of Educational Research and Training, for example, produces national and state-level assessment reports for the National Student Assessment (NAS) as well as summary reports [3]. Producing multiple reports has the advantage that the reports can be more targeted to stakeholders and to priority areas and by producing shorter reports they can be published in a timelier manner.

Agreement should be reached amongst the assessment team and stakeholders (see Chapter 2) about what reports should be written and the structure of these reports. Developing reports is generally an expensive and time consuming process and, therefore, priorities need to be agreed upon.

Some general principles when writing and disseminating reports are provided in Exhibit 48.

**Exhibit 48:** Report writing considerations

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report results accurately</td>
<td>Ensure results are reported accurately and that any limitations and uncertainties are acknowledged. Include sufficient detail so that readers have a full picture of the results.</td>
</tr>
<tr>
<td>Be concise</td>
<td>Balance providing sufficient detail with writing concisely, so that reports are easily accessible to readers and so that the essential information stands out.</td>
</tr>
<tr>
<td>Present information in appropriate ways</td>
<td>Ensure that appropriate tables and graphs are used to present data, for example, line graphs may be used to indicate trends over time. Provide explanations of tables and graphs for readers.</td>
</tr>
<tr>
<td>Use clear and simple language</td>
<td>Reports are best when written in the active rather than passive voice; when they use easily understandable language; and when technical terms are explained (in non-technical reports). Take into account the likely background knowledge of the target audience.</td>
</tr>
<tr>
<td>Ensure the report is well structured and of high quality</td>
<td>First impressions count. Readers may make judgements about the quality of the assessment based on the quality of the report, or they may not continue reading if it is not presented in a logical and professional way. In particular, ensure logical sequencing; use a style guide; and gather feedback on draft reports.</td>
</tr>
<tr>
<td>Make reports easily accessible and available in a timely manner</td>
<td>If reports are to be of interest to stakeholders, they need to address current policy priorities. Therefore, they need to be produced in a timely manner. Where possible, make reports publicly available and easily accessible.</td>
</tr>
<tr>
<td>Ensure the report is of interest to the target audience</td>
<td>Write in a way that engages readers, so that they will want to find out more. Ensure that the report addresses questions and policy issues that will be of interest to the target audience. For example, if the target audience is teachers, the report may include suggestions for teaching based on the assessment results.</td>
</tr>
</tbody>
</table>
14.3.1 Technical report

The technical report is likely to have a smaller and more technically knowledgeable audience than the other types of reports. Technical reports are written for most high-quality large-scale assessment projects as they provide the information necessary to interpret the results and information for the audience to make judgements about the robustness of different aspects of the assessment. If stakeholders judge assessment projects to be of high-quality, then this will increase the likelihood of the assessment data being utilised [4].

A technical report generally includes the following sections:

- assessment purpose
- what is being measured, including domain definitions
- instrument development
- translation
- population or sample assessed
- field operations
- scoring
- data entry and cleaning
- item analysis
- item scaling
- analysis of assessment data
- technical standards, including the extent to which they were met
- conclusions.

The technical report also provides a valuable record for the assessment team of the technical aspects of the assessment program. This information can be used when implementing future assessment cycles, in order to ensure that the same procedures are used, or where change is desired, that these changes are implemented deliberately. Having this record can be particularly useful if there has been staff turnover between assessment cycles, and staff members are no longer involved.

14.3.2 Main report

The contents and structure of the main report will vary for different learning assessments. A main report typically includes the following information:

- context and objectives of the assessment
- assessment framework and instruments
- procedures followed, including sampling, field operations, data analysis and limitations
- descriptions of student achievement, including differences between subgroups (example items may also be included to illustrate students’ proficiency levels)
- correlates of achievement
- changes in achievement over time, where this data is available
- overall summary of findings, recommendations and conclusion.

The previous chapters in this handbook have discussed many of these different areas: context and objectives of the assessment (Chapter 2), assessment framework and instruments (Chapters 4-8), procedures followed (Chapters 9-12), assessment results – descriptions of student achievement, correlates of achievement, changes over time (Chapter 13). Three important aspects – the limitations, summary of findings and recommendations – which have not already been discussed in this handbook are now briefly outlined.
14.3.2.1 Acknowledging assessment limitations

Every assessment has limitations, and following principles of good assessment practice, as discussed in previous chapters, will help to minimise the limitations. It is important that any limitations of the learning assessment are acknowledged.

Limitations of the study may be reported in a separate section of the report, or they may be incorporated into the relevant sections. For example, any sampling issues may be discussed in the sampling section of the report.

The results should be presented in a way that takes into account the assessment limitations. For example, differences in achievement between different districts should not be presented if this analysis was not taken into account during the design of the sample.

Acknowledging the limitations can also help minimise over-simplistic interpretations of the results. For example, if the media reported that higher achievement in a state was the result of smaller class sizes, this would likely be an over-simplistic interpretation. There are many contextual factors associated with achievement, and in addition, it is difficult to establish causality. The assessment report needs to acknowledge the complexities and uncertainties around influences on student achievement so that the results can be interpreted and used appropriately.

14.3.2.2 Presenting a summary of findings

In order to draw readers’ attention to the most important information, a summary of findings section is often included at the start or end of assessment reports. To identify the key findings, consider the following questions: What is the information that is essential for people to know? What is the information that people want to know? What information could be used to make improvements to teaching and learning? What information could help inform the educational policy goals?

The summary of findings includes the essential information in brief, often in the form of a list of bullet points. Of course, it is important that it is clear to readers where they can find more detailed information about these findings, which can be achieved through clear report subheadings and a logical report layout.

In addition to the summary of findings section, it can also be helpful to include a short bullet-point list of the key information at the start of each chapter. For example, the descriptions of student achievement chapter may start with three or four points to be highlighted. Reiterating the core messages increases the likelihood that audiences will read and remember the results.
14.3.2.3 Making recommendations based on the assessment

Some assessment reports include a set of recommendations about possible next steps. These recommendations may be based on the assessment data alone, or may also draw on findings from other assessment, evaluation or research findings. Recommendations may also include suggestions for further research or for future assessment phases.

Other assessment reports still make links to policies and practices but do not explicitly make recommendations. When deciding whether to include recommendations in assessment reports, considerations include:

- the project stakeholders’ expectations about whether these should be included – this should be clarified early on in the project
- the assessment team’s ability to formulate useful recommendations based on the assessment data – for example, does the team have a good understanding of the context?
- the opportunities stakeholders have had to be involved in discussing the assessment data – recommendations which have been discussed with stakeholders who have in-depth knowledge of the sector are likely to be more robust
- the technical quality of the learning assessment, and therefore, the validity of the recommendations – for example, if the assessment instrument is much too easy for most students, this will have an impact on how useful the assessment data is.

As will be discussed in the next section, if the assessment results and recommendations are to have an impact on policy, this involves much more than just reporting. Policy responses are rarely obvious or clear-cut, and it is therefore essential that possible next steps are discussed and debated with the assessment team and stakeholders. Some of the facilitating factors to assessments having an impact on policy are outlined below.
14.4 Facilitating policy impacts

In order for learning assessments to have an impact on policies, the assessment team may need to work with senior bureaucrats and government ministers to unpack the assessment results and to discuss the feasibility and consequences of various policy options.

One common issue for learning assessments is that after the reporting and dissemination, the results can have little ongoing impact. Policy decisions are complex, drawing on many sources of information over time, and therefore, it is sometimes difficult to identify the impact of learning assessments on policy. However, there are many factors that influence the likelihood of assessment results being used by stakeholders. This section will focus on those factors over which the assessment team can have some control.

14.4.1 Assessment quality

The technical robustness of the learning assessment is one of the key factors that contribute to assessment data being used by policymakers [4]. Stakeholders need to be confident that they can trust the results and interpretations. Following the best practices outlined throughout the different chapters in this handbook will help ensure the quality of the assessment. As discussed earlier in this chapter, providing stakeholders with sufficient information about the procedures followed – for example, in the form of a technical report and responding to stakeholder questions – will provide evidence about the quality.

In order to conduct a high-quality assessment, the assessment team requires a diverse range of skills and knowledge. Assessment teams may benefit from partnering with other organisations which can fill any gaps in the team’s expertise.

14.4.2 Capacity of stakeholders to use information

Stakeholders may vary in their ability to understand assessment data and their ability to apply the results. The assessment team can play a role in building this capacity by involving stakeholders throughout the assessment program. The results should be explained to stakeholders in a way that shows their relevance and makes them easily understandable. Holding workshops or conferences where the results are discussed can be particularly effective, as this allows stakeholders to ask questions and provides opportunities for the assessment team to work with stakeholders to unpack the results and possible implications. It is important that stakeholders are supported to make appropriate use of the data and to ensure they have realistic expectations about what information the assessment can and cannot provide.

If possible, the assessment team should bring together the stakeholders with the right mix of skills, knowledge and influence. Policymaking is a complex process often involving many different organisations. Through organising meetings and workshops, the assessment team can ensure that the appropriate decision-makers are involved. Stakeholder discussions can be particularly valuable when the results are politically sensitive and where diplomacy around formulating recommendations is crucial.
14.4.3 Connection to policy processes and other educational activities

Learning assessment results are also more likely to be utilised when they are integrated into the policymaking and decision-making processes of the country or state [4]. For example, in some countries this is achieved by a well-respected assessment body being set up within the ministry of education, while other countries ensure that there are feedback loops between the assessment agency and government [4]. Systems and strategies need to be in place so that the learning assessment findings reach the decision-makers at appropriate times in the policy cycle.

Learning assessments are more effective when they are connected with other educational activities, rather than conducted in isolation. Drawing on multiple pieces of evidence, including from assessments, research and evaluation projects, can strengthen the support for making policy decisions. Caution is needed to ensure that the evidence used is relevant and of high quality.
14.5 Policy impacts

Policymakers are typically the primary audience for large-scale assessments, as they are the most likely to be able to enact change, which can, in turn influence other stakeholders, such as teachers, school leaders and students. For example, learning assessments can be used to:

- understand achievement levels, in order to identify policies to improve achievement for the population as a whole or for subgroups. For example, if overall achievement in writing was lower than expected, increased resources may be required to improve student learning outcomes in this domain.
- understand the resources available and identify where resources are needed. If contextual data is collected, this could show, for example, states or districts in India where there are a lack of trained teachers.
- monitor achievement levels over time to identify whether education system changes have had an impact. This can then be used to make decisions about whether the education system changes are effective and should be continued, or whether a different approach should be taken.
- set standards, or achievement levels for what students should be able to do at certain stages of their schooling. It is important that these standards are realistic.
- make curriculum and/or textbook changes, for example, through gathering information about how the curriculum is being implemented.
- identify pre-service and/or in-service professional development needs of teachers. For example, the National Achievement Survey (NAS) [5] collects contextual information on a range of factors, such as teachers’ attendance at teacher training programs, teachers’ planning and support and use of teaching materials. This can be used to make decisions about pre- and/or in-service training gaps.

Policy decisions should take into account understandings of best practice and be economically viable. If interventions are to be successful, organisations, schools and other stakeholders need to have sufficient resources and support to enact change.

Assessment teams should keep track of the different ways that assessment data is utilised. For example, they can identify when assessment data has been used in research papers and in the media, and how the reporting has informed education policies and programs.

Assessing the longer-term impacts is also essential in order to evaluate whether the policies are having the desired outcomes. This could be in the form of an evaluation and/or through future learning assessment phases. For example, the NAS found that ‘between 2000 and 2011, the Government of India increased its financial commitment to elementary education six-fold; yet learning levels did not improve commensurately’ [6, p. 357]. Assessments such as this can provide valuable information to policymakers over time, so that policy decisions can be evidence-based.
14.6 Summary

Learning assessments have the potential to have wide-reaching impacts on educational policy which can, in turn, benefit teaching and learning. As this chapter has discussed, in order for learning assessments to make a positive impact, the reporting, dissemination and consultation activities need to be well designed and purposeful. Stakeholders, and in particular policymakers, are more likely to use data from an assessment to inform educational policy when the assessment prioritises their policy concerns, is technically robust, and produces reliable and valid data and ensuing analyses.
# Checklist: Reporting and dissemination

If you are planning or implementing a large-scale assessment, have these points been addressed?

## Planning the dissemination strategy

1. Have the dissemination activities been planned and budgeted for at the beginning of the assessment program?  
2. Has the dissemination plan been revisited throughout the assessment program, and adapted where necessary?  
3. Have the key groups and individuals that have an interest in the assessment been identified (the target audiences)?  
4. Have the needs of the target audiences been identified?

## Selecting the dissemination methods

5. Have different dissemination methods been considered and evaluated for their effectiveness?  
6. Will the dissemination methods selected communicate the results to a wide range of stakeholders?  
7. Are the dissemination methods selected likely to help stakeholders understand and use the results?

## Developing reports and other products

8. Has a high-quality report detailing the technical aspects of the project been written?  
9. Has a high-quality results report been written?  
10. Have other high-quality products been developed (e.g. interim reports, thematic reports, pamphlets, media releases, websites and newsletters)?

## Engaging stakeholders

11. Have steps been taken to build the capacity of stakeholders to use the assessment findings and to inform policies?  
12. Have stakeholders had opportunities to discuss and debate the implications of the assessment findings?

## Facilitating policy impacts

13. Have quality assurance methods been used throughout the assessment program?  
14. Have the assessment results been integrated into the policymaking process?  
15. Is the assessment connected to other educational activities?

## Documenting and evaluating policy impacts

16. Have any policy impacts of the assessment been documented?  
17. Are there plans to assess the longer-term impacts of any policy changes as a result of the assessment?
**Glossary**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissemination</td>
<td>The sharing of information.</td>
</tr>
<tr>
<td>Domain</td>
<td>The area of learning that is the focus of an assessment. This may be a curriculum area (e.g., mathematics or science), or a more generic area of learning (e.g., reading, writing or problem-solving).</td>
</tr>
<tr>
<td>High-stakes assessment</td>
<td>An assessment is considered high stakes when there are serious consequences connected to the outcomes. For example, an entry examination for a university is high stakes.</td>
</tr>
<tr>
<td>Items</td>
<td>The questions or tasks used in an assessment.</td>
</tr>
<tr>
<td>Stakeholders</td>
<td>People/groups that have an interest in the assessment/results.</td>
</tr>
<tr>
<td>Target audience</td>
<td>A particular group that the information/report/assessment results are aimed at.</td>
</tr>
<tr>
<td>Webinar</td>
<td>A seminar facilitated via the internet. Often called web conferencing and similar to video conferencing.</td>
</tr>
</tbody>
</table>
References


5. NCERT. (2012). National Achievement Survey Class V. New Delhi: NCERT.

Contributors

The following were involved in the development of this handbook, including writing, reviewing and editing.

ACER

The handbook chapter authors were:

Abha Bhagat (Chapter 6)
Tim Friedman (Chapters 5 and 8)
Jennifer Hong (Chapters 9 and 10)
Renee Kwong (Chapters 12 and 13)
Juliette Mendelovits (Chapters 5 and 6)
Louise Ockwell (Chapter 8)
Sally Robertson (Chapters 1 and 14)
Bikramjit Sen (Chapter 6)
Stephanie Templeton (Chapter 7)
Alvin Vista (Chapters 4 and 11)
Maurice Walker (Chapters 2 and 3)

The handbook editors were Sally Robertson, Alvin Vista and Maurice Walker, and the proof reader was Rebecca Leach. Rudra Sahoo participated in the development of the structure and coverage.

NCERT

Prof. Y. Sreekanth, Professor & Head, ESD, NCERT
Prof. Sridhar Srivastava, Professor, ESD, NCERT
Dr. A.D. Tewari, Associate Professor, ESD, NCERT
Dr. Santosh Kumar, Associate Professor, ESD, NCERT
Dr. Veer Pal Singh, Associate Professor, ESD, NCERT
Dr. Indrani Bhaduri, Associate Professor, ESD, NCERT
Dr. Sukhvinder, Assistant Professor, ESD, NCERT
Dr. Vishal D. Pajankar, Assistant Professor, ESD, NCERT
Dr. Satya Bhushan, Assistant Professor, ESD, NCERT
Dr. Gulam, Assistant Professor, ESD, NCERT
Dr. Manika Sharma, Consultant, ESD, NCERT

State education bodies

Dr. G. Narasimha Reddy, Lecturer, SCERT, Andhra Pradesh
Dr. Richard Cabral, Coordinator, SCERT, Goa.
Mr. Manoj Koradia, Research Associate, SCERT, Gujarat
Mr. Haresh Chaudhari, Curriculum Coordinator, SCERT, Gujarat
Mr. Vitthal K.Gorde, Lecturer, SCERT, Maharashtra
Mr. Daniel Thong Seb, Training cum Research Assistant, SCERT, Nagaland
Mr. Suraj Bir Singh, Coordinator Maths, SIE, Sikkim
Dr. Nandan S. Bisht, Research Officer, Uttarakhand Board of Secondary Education
Ms. M. V. Kumari, Senior Assistant Director, DSERT, Karnataka

RMSA-TCA

Dr. Jayshree Oza, Team Leader, RMSA-TCA
Parul Pandya, Research and Technical Support Officer, RMSA-TCA
Jason Pennells, Education Adviser, Cambridge Education
Daniel Harris, Communications Technical Adviser, RMSA-TCA
Supported by

RMSA - Technical Cooperation Agency
Room No. 308 - 313
Central Institute of Educational Technology (CIET) Building
National Council of Educational Research and Training (NCERT)
Sri Aurobindo Marg, New Delhi-110 016
T  +91-11-26569925 / 26560559
E  contact@rmsatca.org

The RMSA Technical Cooperation Agency is funded by the UK Department of International Development (DFID)