Test development.
Part I
Practical Foundations
The purpose of this chapter is to explain how the psychometric principles outlined in the remaining chapters of this Handbook can be applied in order to develop a test. We take a broad definition of what constitutes a test and what is understood by test development. This is because the principles of psychometric testing are very broad in their potential application. Among others, they can apply to attitude, personality, cognitive ability, interest, and diagnostic measures. For the purposes of this chapter all such measures will be referred to as tests. Psychometrics is broad in another sense: It applies to many more fields than psychology; indeed, biomedical science, education, economics, communications theory, marketing, sociology, politics, business, and epidemiology, among other disciplines, not only employ psychometric testing, but also have made important contributions to the subject. Our definition of a test is broad in another sense: It encompasses everything from a simple attitude scale, say to measure job satisfaction, to comprehensive test batteries such as the Woodcock–Johnson IV battery of cognitive tests (Schrank, Mather, & McGrew, 2014). Of course, not every aspect of test development applies to both, but the overlap is considerable.

It may be useful to distinguish the different levels of complexity involved in test development. In the simplest case, the test comprises just one scale, but more usually a test is comprised of multiple scales (single scale versus test battery). A second distinction is between tests comprised of similar as opposed mixed types of scales (scale similarity). For example, the European Social Survey measures multiple constructs but all are attitude scales. However, some instruments may combine assessments of mixed scale types; for example, cognitive ability, personality, and attitudes. A third dimension concerns whether the test is intended to sample the entire spectrum of a domain, or whether it is focused on specific aspects (broad versus narrow spectrum). For example, it would not be feasible for a selection test to reliably measure all facets of either personality or cognitive ability. The point being that some form of systematic choice procedure is required such as job analysis or meta-analysis (Hughes & Batey, 2017). Fourth, there is the issue of team size. There is a very big difference from the situation in which a single investigator takes responsibility for the major
portion of test development, and the situation in which there is a large team with diverse skill sets, which would be common when developing commercial tests. The MAT\textsuperscript{80} (Irwing, Phillips, & Walters, 2016), which we use later to demonstrate test development procedures, is a test battery with a mixed scale that combines personality and ability scales, involved a small test development team, and requires systematic selection of specific facets.

There are already many publications of relevance to the topic of test development. Probably the most useful single source is “The Standards for Educational and Psychological Testing” (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 2014). However, as its name implies, this tells you what needs to be done, but not how to do it. There is now a very useful *Handbook of Test Development* (Downing & Halady, 2006), which largely specializes in the design of educational and ability tests. Of almost equal use are textbooks on questionnaire and survey design (Brace, 2005; De Vaus, 2014; Foddy, 1996; Oppenheim, 1992). Perhaps what none of these books quite do is link modern psychometrics to test development, which is the aim of this chapter and the whole Handbook.

We begin with a comprehensive model of the stages of test development, and then discuss the major considerations that apply at each stage. We will leave the reader to decide which of these stages apply to their own situation, depending on the type and purpose of the test. Table 1.1 outlines a 10-stage model of test development. There are a number of stage models of test development in existence (e.g., Althouse, n.d.; Downing, 2006) and, to a degree, such models are arbitrary in the sense that which tasks are grouped into a stage and the order of stages is probably more for explanatory convenience rather than a description of reality. In practice, tasks may actually be grouped and undertaken in many different combinations and orders, with many tasks undertaken iteratively. Nevertheless, a stage model provides a systematic framework in which to discuss the tasks that must be undertaken, although not all tasks are relevant to all types of test development.

**Table 1.1** Stages of test development.

<table>
<thead>
<tr>
<th>Stages and substages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Construct definition, specification of test need, test structure.</td>
</tr>
<tr>
<td>2 Overall planning.</td>
</tr>
<tr>
<td>3 Item development.</td>
</tr>
<tr>
<td>a. Construct definition.</td>
</tr>
<tr>
<td>b. Item generation: theory versus sampling.</td>
</tr>
<tr>
<td>c. Item review.</td>
</tr>
<tr>
<td>d. Piloting of items.</td>
</tr>
<tr>
<td>4 Scale construction – factor analysis and Item Response Theory (IRT).</td>
</tr>
<tr>
<td>5 Reliability.</td>
</tr>
<tr>
<td>6 Validation.</td>
</tr>
<tr>
<td>7 Test scoring and norming.</td>
</tr>
<tr>
<td>8 Test specification.</td>
</tr>
<tr>
<td>9 Implementation and testing.</td>
</tr>
</tbody>
</table>
Construct Definition, Specification of Test Need, and Structure

The motivation for test development often stems from a practical concern: can we help children learn, can we identify effective managers, can we identify those at risk of mental distress? However, while motivation may provide impetus, it is not the formal starting point of test development. The formal starting point for all test development is to generate a construct definition, which broadly is a definition of what is to be measured. An initial construct definition should be as clear as possible but will often be somewhat broad. For example, one might decide that a measure of cognitive ability, or leader potential, or anxiety is required (perhaps in order to address our previous practical concerns). From this point, one can define these constructs (using extant work or a newly generated definition as appropriate) and conduct a systematic literature review to identify existing tests and find out more about the nature of the target construct. This review should help the developer to refine their construct definition. For example, if you were interested in measuring cognitive ability an initial definition might be incomplete or very high level (e.g., ability to acquire and use knowledge or the speed and accuracy of information processing). However, based on a literature review, one could choose to devise sufficient tests to provide coverage of all second-order factors of cognitive ability contained within the Cattell–Horn–Carroll model (McGrew, 2009). This was broadly the strategy used in the development of the Woodcock–Johnson IV (McGrew, 2009).

However, relying solely on extant models might not always be the most useful strategy for at least two reasons, which we will explore using the Five Factor Model (FFM) of personality (Costa & McCrae, 1995). First, because the FFM is so widely accepted, there already exist a large number of tests based on this model and the question then arises as to what is the need for another identical test. Second, although the FFM is widely accepted, it seems unlikely that it is the final word on personality. Some argue that there are facets of personality out with the sphere of the FFM (Paunonen & Jackson, 2000), including some aspects of abnormal personality (Mathieu, Hare, Jones, Babiak, & Neumann, 2013). Of course, the NEO-PRI was not designed to measure abnormal personality, but there are strong arguments that broad-spectrum measures of personality should cover both the normal and abnormal (Markon, Krueger, & Watson, 2005). This may seem like a disadvantage but of course, from the point of view of a test developer, it is an opportunity. There is much more value in a new test that does something that an old test does not.

There may of course be many reasons for developing a test. There may be a need for research on a topic, but no extant measure suitable to carry out the research. For example, knowledge is a very important aspect of human behavior, but until about the year 2000 there were no standardized tests of knowledge (Irving, Cammock, & Lynn, 2001; Rolfhus & Ackermann, 1999). Outside of research: diagnosis, assessment, and development, employee selection, market research, licensing and credentialing (e.g., the examinations that qualify one to practice as an accountant or lawyer) represent other broad categories of test needs. Broadly, there is a need for a test if your systematic literature review reveals that a test does not currently exist, current tests are judged to be inadequate, or there are tests, but not ones suitable for the particular population or use to which the test is to be put. Certainly, many instances of copycat tests exist, but I am not aware that this strategy has generally proven to be a recipe for a successful test.
Generally, successful tests are developed due to some combination of three circumstances:

1. Theoretical advances (NEO PI-R: Costa & McCrae, 1995; 16 PF: Conn & Rieke, 1994; VPI: Holland, Blakeney, Matteson, & Schnitzen, 1974; WAIS: Wechsler, 1981);
2. Empirical advances (MMPI: Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989);
3. A practical (market) need (SAT: Coyle & Pillow, 2008; GMAT: Oh, Schmidt, Shaffer, & Le, 2008).

If the developer does not make a test based on theoretical advance, empirical advance, or a gap in the market and instead duplicates a test, or more realistically produces a test that shares a name with another but has subtle differences in content, then the result is construct proliferation and the well-documented problems commonly referred to as the Jingle-Jangle fallacy (Hughes, Chapter 24; Shaffer, DeGeest, & Li, 2016).

Theoretical and empirical advances

Theoretical advancements (often driven by prior empirical discoveries) undoubtedly provide the reason for the development of many tests. Briefly, the test developer must develop a theoretical framework, which is in some respect new and sounder than previous frameworks, or utilize existing theoretical frameworks that current tests have not exploited. A full discussion of the nature of theoretical advances is well beyond the practical bounds of this chapter because it will be unique for every construct. That said, the history of the development of the FFM is highly instructive as to the process whereby theory evolves from an interaction between theoretical and empirical developments (Block, 1995; John, & Srivastava, 1999, see later). Also, pivotal to test development is the evolution of tight construct definitions, which also emerges from the interaction between theory and empirical work.

Systematic domain mapping

Perhaps the most obvious example of an interaction between theoretical and empirical advance comes in the form of systematic domain mapping. Very simply, a systematic domain map consists of all construct-relevant content (e.g., every aspect of the domain of personality) mapped onto a theoretically supported structure. This serves as a precursor to developing a systematic taxonomy of the domain that ideally identifies all primary level and higher-level constructs and provides the basic material from which test items can be constructed.

The history of testing suggests ways in which this can be achieved. Although all attempts to map a domain suffer from practical and statistical limitations. For example, the total number of possible personality items is sizable and collecting data on so many items is difficult as is subsequent analysis. For instance, factor analysis cannot handle the size of data matrix that would be required, meaning that in practice the total domain needs to be divided into manageable chunks based on a subjective grouping (see Booth & Murray, Chapter 29). The process of grouping items inevitably means that
some constructs which span the subjective groupings or sit at the interface between two are not sufficiently captured. Nevertheless, the development of the FFM, for example, is instructive both as to how domain mapping can be achieved and also the potential flaws in this process. Actually, the history of the development of the FFM is complex (Block, 1995; John, Angleiter, & Ostendorf, 1988), but a simplified account of the principles of its development will suffice for our purposes. Arguably, the development of the FFM stems from the lexical hypothesis, which is comprised of two major postulates. The first states that those personality characteristics that are most important in peoples’ lives will eventually become a part of their language. The second follows from the first, stating that more important personality characteristics are more likely to be encoded into language as a single word (John et al., 1988). If true, then in principle, if all words that describe personality were incorporated into a questionnaire, and a large population were rated as to the extent these words apply to them, then a factor analysis of this data would provide the facet and higher-order structure of personality. In practice, despite claims to the contrary, for various practical reasons this has never been done, but something like it has (e.g., Ashton, Lee, & Goldberg, 2004). Personality research is now at a stage at which there are many respected measures of personality and the next step might be to administer all known measures of personality to a large population and, guided by theoretical developments, factor analyze the resultant data set in order to provide a new and more comprehensive taxonomy of personality (Booth, 2011; Woods & Anderson, 2016).

What this example illustrates is that successful test development often requires some form of systematic domain mapping, which is in at least some respects novel.

Practical (market) need

Of course, measures derived from a taxonomy or theory do not necessarily correspond to a practical need (beyond the need for measurement). Indeed, one difficulty with omnibus measures (such as the Woodcock–Johnson IV and the NEO PI-R) is that they rarely correspond to a direct market need. In the most part, this is because omnibus measures are often long and time-consuming to complete, resulting in equally long and detailed reports. Exhaustive reports concerned with all aspects of personality or cognitive ability can be difficult for laypersons to understand and use. Usually, the tests adopted by consumers are shorter and considered more user friendly. For example, despite being technically deficient (Hughes & Batey, 2017), the MBTI is among the most commonly used personality tests because it is relatively short, the results are easily communicated and understood, and therefore it can readily be used in a practical context. Probably therefore, marketable tests may be based on a systematic taxonomy but the actual constitution of the test will depend on additional considerations. In short, for a test to address a market need it should be both technically sound (in terms of theoretical grounding and psychometric properties) and practically useful.

The area of selection can help illustrate what some of these additional practical considerations might be. One starting point might be to identify the market for a selection test based on systematic market research. Let us imagine that the results of this research reveal there to be a large market for the recruitment of managers, not least because a large number of managers are employed, and secondly because their characteristics are often considered crucial to the success or failure of companies. How then could
we devise a test for managers? Traditionally, most test developers for a selection instrument would begin with a job analysis (Smith & Smith, 2005). This is still an essential step in the development of selection tests, however, since the late 1970s psychometric meta-analysis has become an important source of information to guide the development of selection instruments.

Meta-analysis

The main purpose of psychometric meta-analysis is to obtain parameter estimates, which are unbiased and corrected for measurement artifacts. Hunter and Schmidt (2004) is probably the most useful introduction to meta-analysis, although some more recent developments are contained in Borenstein, Hedges, Higgins, and Rothstein (2009). Meta-analysis has many potential applications to test development. For example, with regard to the construction of test batteries for employee selection, findings of meta-analyses identify which constructs predict future job performance and, therefore, which should be included (e.g., Judge, Rodell, Kliner, Simon, & Crawford, 2013; Schmidt, Shaffer, & Oh, 2008).

Psychometric meta-analysis averages the value of an effect size across studies in order to obtain a reliable summary estimate. The most important effect size in a selection context is the predictive validity, which is measured by the correlation between the score on the selection measure and some measure of job performance. The biggest problem with most estimates of predictive validity from single studies arises from sampling error, which is more considerable than is generally imagined. As sample size tends to infinity, so sampling error tends to zero and thus by amalgamating findings across studies, large meta-analyses effectively reduce sampling error to miniscule proportions. Standardly, psychometric meta-analysis also corrects for artifacts due to error of measurement, range restriction, imperfect construct validity (e.g., different measures of purportedly the same personality construct typically correlate at 0.4–0.6, see Pace & Brannick, 2010), use of categorical measurement, study quality, and publication bias. However, once these corrections are made, the confidence interval around the effect size estimate may still be large. This may indicate that the effect size is dependent on a third variable, usually referred to as a moderator. For example, cognitive ability predicts more strongly for complex jobs (Schmidt & Hunter, 1998) and in the case of personality, traits predict more strongly when they are relevant (e.g., Extraversion and sales; Hughes & Batey, 2017).

The findings of meta-analysis with regard to which cognitive abilities and FFM personality factors predict job performance are, within limits, fairly definitive (Schmidt & Hunter, 1998, 1998; Schmidt, Shaffer, & Oh, 2008). Virtually every meta-analysis that has investigated the issue has concluded that, for most jobs, general cognitive ability is the best predictor and the level of prediction increases in proportion to the cognitive demands of the job (Schmidt & Hunter, 1998). Moreover, it is generally contended that second-order factors of cognitive ability such as spatial, verbal, and memory add little incremental prediction (e.g., Carretta & Ree, 2000; Ree, Earles, & Teachout, 1994). Although it is a hotly contested issue, meta-analyses of the predictive validity of personality show virtually the opposite; that is, that personality largely does not offer blanket prediction of job performance across roles. Some have argued from this data that personality tests should not be used in selection (Morgeson et al., 2007), but many have also argued otherwise (e.g., Ones, Dilchert,
Viswesvaran, & Judge, 2007; see Hughes & Batey, 2017, for a comprehensive review). So you would think that any selection program for a management position would incorporate a measure of cognitive ability. Actually, this is not so. In practice, it would seem that personality tests are more widely used in selection than cognitive ability tests; perhaps because of greater acceptability.

Of course, meta-analysis has its limits. Meta-analysis can only be applied where there is already a substantial body of research. For example, because of problems of acceptability there is a huge paucity of work on dark traits of personality, yet what work there is suggests that it is indeed important for organizational performance (Babiak, Craig, Neumann, & Hare, 2010; Boddy, 2011, 2014). To summarize, meta-analysis provides a very useful body of evidence about what works, but this needs to be balanced against what is acceptable. Whatever, decision is made, the findings of meta-analyses place the development of test batteries for selection on a sounder basis than has previously obtained.

We have considered some scenarios, and we will consider others in the context of item development, however, the range of possible scenarios for test development is large and you should consult other sources for aptitude and achievement tests, for example (Downing & Haladyna, 2006). Whatever the exact scenario, at the end of this stage you will have defined and refined your constructs, specified the test need (including the purpose of the test and the populations to which it will be administered), and identified the exact components to be included within your test.

**Overall Planning**

Once the purpose and test structure is defined, a process of planning is required that may be more or less extensive depending on whether the test involves just a single research scale or is designed to measure a large number of different constructs.

The planning phase involves answering a broad range of questions. For example:

1. How many items are needed to measure the constructs of interest?
2. Which response scale is to be used? A basic choice is between multiple choice formats, Likert type scales,\(^1\) and constructed formats in which the test taker supplies a response to an open question.
3. How to score the test? For different purposes, it may be appropriate to use sum scores, standardized scores (e.g., t scores, Stanines, Stens, or IQ values), or some form of item response model for scoring.
4. Which psychometric model is most appropriate for modeling the test data? Usually the choice is posed as between a classical test theory approach and one based on IRT, however, we will suggest next that perhaps a combination of the two is to be preferred.
5. What item development process is most suitable? Item trialing and validation studies also have to be planned, together with the analyses of the resulting data. In addition, quality control systems need to be developed or adopted in order to ensure; for

\(^1\) Strictly, a Likert scale is a summative scale in which the total score is the sum of the item scores. However, the term is commonly used to refer to a scale of a type with verbal anchors, most usually ranging from strongly agree to strong disagree. This is the sense intended here.
example, that the test specification is completely accurate and it is implemented precisely (see the later section on Test Specification).

6 How to administer the test? The modality of administration could be paper and pencil, interview, telephone interview, or computer based. There is no single format that is “best,” but computer-based administration is becoming the preferred option since for many tests as it has the advantages of accessibility (anyone anywhere in the world who has a computer can take the test), automatic data capture, and instantaneously delivered feedback. The principal disadvantage of computer-based administration is that it may not provide access to segments of the population who do not have computer access. There is also the question of how to prevent cheating.

Once these decisions have been made a timeline is required that should include time for correction cycles, which will almost certainly be necessary. This should specify who is responsible for each major task together with completion dates (Roid, 2006). Lack of specificity of the plan will undoubtedly delay development and cost money.

Item Development

The nature of item development depends on the type of scale to be developed. The type of scale depends on what is to be measured, the response format, and scaling model employed. Broadly, there are attitude, trait, and ability scales. The most commonly employed response formats are Likert type, multiple choice, or forced-choice items. Scaling generally conforms to the types developed by Thurstone, Likert, or Guttman (Ghiselli, Campbell, & Zedeck, 1981). To a degree, the type of scale, response format, and approach to scaling has an effect on recommendations as to item writing and development.

Irrespective of the type of item to be developed, the first stage consists of the development of further refined construct definitions. This is an issue that is typically revisited several times. For some cognitive abilities, the construct is already extremely well defined. A good example is that of matrix reasoning tests. Carpenter, Just, and Shell (1990) have defined a set of rules that can be applied in order to develop such tests, and even suggest how different combinations of rules affect item difficulty. In order to apply these rules, all that is required is to define the tokens of which items are to be comprised and how these tokens are to be organized spatially. This clarity of construct definition underpins the development of automatic item generation described by Morrison and Embretson (Chapter 3). Alternatively, provided the construct definition is sufficiently clear, it is quite common practice to use professional item writers.

Achievement tests in contrast to pure measures of ability require mastery of a subject matter. Most tests of this type are developed in conjunction with subject matter experts. Recommendations for the development of such tests are contained in Downing and Haladyna (2006). Downing (2006) in particular specifies a 12-step procedure for the development of such tests.

Trait and attitude items are generally somewhat less well defined, with construct definitions normally arising either directly or indirectly from an exercise in domain mapping. Domain mapping at the test and item level involve similar processes. At the item level, domain mapping techniques can be focused on just those constructs to be measured. This process is so important that there is value in suggesting a number of additional strategies in order to accomplish this.
1 First, define the topic: political attitudes, car buying, environmental issues. Then ask a large and ideally population representative sample what they consider the key issues to be in relation to the topic. At a second stage, an equally large population representative sample can be asked to express what they consider to be commonly held opinions with respect to each issue. This should provide a very long list of statements from which items can be developed. These items should be grouped into scales, which measure just one dimension of the construct in question.

2 Use already extant systematic sources. Examples would include encyclopedias (knowledge tests), curriculum specification (educational tests), dictionaries (personality), diagnostic manuals (DSM-V), and web portals.

3 Use expert informants in order to aggregate comprehensive lists of attitudes, characteristics, successful traits, and critical incidents.

These processes should provide the raw materials for items. There are item writing guidelines for multiple choice items, and for Likert type items. Table 1.2 presents a distillation of rules concerning Likert type items. For equivalent sets of guidelines on multiple choice items you could consult Haladyna, Downing, and Rodriguez (2002).

<table>
<thead>
<tr>
<th>Table 1.2</th>
<th>Item writing guidelines for Likert type items.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guideline</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Each item should tap a different aspect of the domain.</td>
</tr>
<tr>
<td>2</td>
<td>Simple language is preferred. There exist comprehensive lists of words in order of their frequency of usage, at least for the English language (e.g., <em>Corpus of Contemporary English</em>, 2016). In general, more commonly used words are more understandable.</td>
</tr>
<tr>
<td>3</td>
<td>Short items are preferable to long items. However, items should not be shortened so that specificity of meaning is lost. For example, short items comprised of several or more words rather than single words provide more reliable measures of personality.</td>
</tr>
<tr>
<td>4</td>
<td>Items should not be double-barreled. That is items should relate to one and not more than one subject.</td>
</tr>
<tr>
<td>5</td>
<td>Items should be positively phrased. Use of single or double negatives reduces comprehensibility.</td>
</tr>
<tr>
<td>6</td>
<td>Items should not be leading or elicit a prestige bias.</td>
</tr>
<tr>
<td>7</td>
<td>Items should use good grammar, punctuation, and spelling in a consistent manner.</td>
</tr>
<tr>
<td>8</td>
<td>Syntax and grammar should be of the simplest form consistent with conveying meaning.</td>
</tr>
<tr>
<td>9</td>
<td>The meaning of items should not be ambiguous.</td>
</tr>
<tr>
<td>10</td>
<td>In general, personalized phrasing is more involving and therefore preferable. However, this may not be appropriate when content is sensitive. In the latter case, an indirect strategy such as asking how you would advise a friend or how other people behave may be appropriate.</td>
</tr>
<tr>
<td>11</td>
<td>Items should not create attitudes, opinions, or responses that do not already exist. Equally, they should not require knowledge the respondent does not possess.</td>
</tr>
<tr>
<td>12</td>
<td>The frame of reference should be clear; e.g., one year, one week, one day.</td>
</tr>
<tr>
<td>13</td>
<td>Items should not overload working memory.</td>
</tr>
<tr>
<td>14</td>
<td>Answering should not require knowledge which is inaccessible, for instance, either because the respondent has never considered the issue, or because it is too taxing on memory.</td>
</tr>
<tr>
<td>15</td>
<td>Items should not be objectionable.</td>
</tr>
<tr>
<td>16</td>
<td>Items should be grammatically consistent with responses.</td>
</tr>
<tr>
<td>17</td>
<td>In general, items should hold the same meaning for all respondents. Following the above rules should help achieve this.</td>
</tr>
</tbody>
</table>
A further issue is that items should not produce response sets which either bias responses or produce response artifacts, although in general this is more a function of the organization and sequencing of items. There are many examples of response sets: yeah-saying, nay-saying, consistency and availability artifacts, halo, and social desirability artifacts. One of the most effective methods of countering response sets is to use forced-choice item formats, the subject of which is discussed extensively in Brown, Chapter 18.

Guideline 8 that, “syntax and grammar should be of the simplest form consistent with conveying the intended meaning,” perhaps requires more comment. Many testees take tests in a second language. There is a large amount of research that shows that reducing the complexity of language also reduces the level of construct irrelevant variance in response to test items (Abedi, 2006). Although much of this research relates to educational and ability tests, it seems highly probable that it applies equally to other types of test. Abedi (2006) provides a useful list of features which reduce linguistic complexity, including: Using commonly occurring and familiar words; using shorter rather than longer words; using shorter sentences; adopting the active voice; using simple rather than compound subjects; reducing the use of comparative phrases, subordinate, conditional, and relative clauses; using concrete rather than abstract language; and using positive phrasing.

Item review

Prior to piloting of items, it is common practice to subject items to a process of review. Probably the most common and arguably useful form of review is the use of expert groups (DeMaio & Landreth, 2004; Presser & Blair, 1994; Willis, Schechter, & Whitaker, 2000). Experts are typically subject matter experts, experts in test design, or representatives of participant samples. The function of different types of experts naturally varies to at least some degree, and of course, some individuals may embody more than one type of expertise. The major concern of subject matter experts is that of item accuracy (the extent to which a test measures what it is intended to measure), but they may also be asked to assess item bias. Very different types of subject matter experts would be required to evaluate items designed to assess personality, mathematics ability, business acumen, knowledge of accountancy, and so on. Experts in test design would be mainly concerned with assessing whether items conform to commonly accepted rules of good item design such as those listed in Table 1.2. In some circumstances, these experts may also be in a position to make accuracy judgments. Experts who are representative of participant samples are primarily used to assess the comprehensibility of the items to the particular population and to identify items that are potentially biased or objectionable. It has become standard practice to eliminate all items flagged as potentially biased or objectionable, unless this would clearly undermine the purpose of the test. Such practices may be misplaced in that there is the risk that informants confuse any form of difference with bias (Messick, 1989). Nevertheless, used well, and obviously dependent on just how expert panels actually are, there is evidence that review by expert groups can be highly effective and relatively inexpensive (DeMaio & Landreth, 2004; Presser & Blair, 1994; Willis et al., 2000).

There are broadly four other methods of item review that are commonly used. Field pretests represent the most commonly used traditional approach. Here the administration of the test is observed in a form as identical as possible to its final form. Debriefings with the observers may be used, the data may be coded to identify problem items, and
even recordings of the test administration may be coded. Cognitive interviews use a combination of thinking aloud protocols and probes. In a thinking aloud protocol the test taker is instructed to verbalize their thought processes as they answer each test item. This may be followed by probes into any source of apparent difficulty in responding to the item. Randomized experiments are also used in which randomly chosen groups are administered different versions of the test in which the wording of items attempting to measure the same thing is varied. Focus groups represent a fourth method in which a semi-structured discussion with members of the target population is intended to explore their knowledge, common opinions, and terms used in relation to the test content. Groves et al. (2009) and Presser et al. (2004) are recommended for more in-depth treatments of these methods.

Piloting of test

To this point, the construction of the test depends on theory, previous empirical work, and subjective judgments based on prior experience and knowledge. The next stages involve administration of the test to an appropriate sample or samples. The sample should closely match the characteristics of the sample or samples in which the test will be used. That said, population representative samples are often useful irrespective of the population(s) at which the test is targeted.

Another consideration with respect to samples is sample size. Assuming that the sampling strategy is constant, bigger is always better in that the larger the sample the smaller the confidence intervals for all parameters. Much ad hoc advice is available on this. For example, 200 is a suggested minimum for factor analysis (Stevens, 2009). In my experience, samples need to be at least 500 to provide generalizable results, and my preference is a minimum sample of 1,000. Kelley and Lai (Chapter 5) describe a systematic procedure for calculating the required sample size for any given preferred confidence interval for factor loadings and other key parameters. This procedure is to be recommended, but ultimately there is a subjective judgment required as to what is an acceptable confidence interval.

Usually tests are administered, analyzed, revised, and readministered a number of times before their psychometric properties are acceptable.

Scale Construction

There is considerable debate about how test data should be analyzed. Here we will make recommendations on what we consider to represent best practice. The issues for initial analyses are whether proposed scales are unidimensional, whether scales actually measure the focal construct (accuracy), how well items measure the underlying construct (reliability), and whether the measure covers the full range of trait values (construct representativeness).

Which analytic techniques are to be recommended depends on the theoretical basis of the test constructs. Nevertheless, in terms of what is currently practicable we propose that in most cases a preferred strategy would involve some combination of confirmatory factor analysis (CFA) and IRT. There are many extensive sources detailing the theory and use of CFA and IRT, but as single sources probably most people would find Brown
CFA is especially useful in terms of providing robust estimates of the number of dimensions underlying a dataset, addressing questions of convergent and discriminant validity, and providing generally optimal estimates of reliability (e.g., McDonald’s Omega, see next). Whereas IRT provides precise information on item difficulty, reliability across the range of scale values, and can provide the basis for short tests with good reliability. For these reasons, we would generally recommend that test development should use a combination of both in order to fully explore the properties of scales and ensure that the final test is optimal in terms of its psychometric properties.

Many textbooks on statistics and articles on methods suggest the use of principal components analysis (e.g., Field, 2013; Hair, Anderson, Tatham, & Black, 2006; Johnson & Wichern, 2007). We suggest that in the majority of cases the disadvantages of using principal components analysis far outweigh the advantages. The choice in terms of theory is between principal components and the common factor model (Mulaik, 2005; Mulaik, Chapter 8). In the first case, principal components analysis is an observed variable model, whereas the common factor model assumes a latent variable model. The difference is very close to the distinction made between cause and effect indicator models (Murray & Booth, Chapter 7). Figure 1.1 shows the difference graphically. In the case of principal components, it is assumed that it is the items that give rise to the score, hence the direction of the arrows from the item to the observed score (see Figure 1.1a). A possible example might be overall health conceived of as the total number of health-related problems suffered in a specified period of time. There is no assumption of one underlying cause to these problems or that there should be a correlation between them, it is just that the sum score provides an indicator of overall health. While this may be an appropriate model for health indicators, for the vast majority of psychometric tests, it is not. Mostly, scientists conceive of their variables as underlying traits that manifest themselves in terms of observed behavior. For example, personality traits are normally conceived of as enduring characteristics of a person, which give rise to consistency in behavior across situations and time (Funder, 2001). They are in effect latent traits that cause manifest behavior. This conception corresponds to the common factor model, which assumes that it is the latent trait that

![Figure 1.1](image-url)  
**Figure 1.1** Principal component model (a) and common factor model (b) represented using structural equation models (SEM) diagram conventions.
causes the observed indicators. Hence, in Figure 1.1(b) the arrows point from the latent variable (not score) to the indicators.

It is often contended that, although the common factor model might be theoretically appropriate, the principal component model has mathematical advantages and that in practice solutions differ only to a minor degree and thus should be preferred. Among these advantages is that principal component is not subject to factor score indeterminacy (Mulaik, 2005) and also many scientists consider principal components to represent the most parsimonious representation of the data. We contend here that these advantages are outweighed by the disadvantages. In the first place, the circumstances in which principal component and common factor solutions are similar are more restrictive than is generally acknowledged; in many cases the differences are quite substantive (Widaman, 1993). To provide examples, it is very rare indeed that a principal component solution will fit a CFA model, whereas exploratory factor solutions using adequate samples and appropriate estimators more often result in good fitting confirmatory solutions (Gerbing & Hamilton, 1996; Hurley et al., 1997). Given the usefulness of CFA as an analytic technique, this is a serious disadvantage. Second, the methods used to estimate the number of factors underlying a data set that are available when using principal components rarely provide the most precise estimate (see Timmerman, Lorenzo-Siva, & Cuelemans, Chapter 11). Third, principal components analysis assumes perfect reliability of measurement, which is highly implausible with respect to most social science and many other types of data. Fourth, however small the difference, why would anyone prefer the theoretically incorrect solution? For these reasons, we suggest that the common factor model in exploratory, but preferably in confirmatory form is to be preferred in most instances.

Whether you can actually use CFA effectively depends on how well the scale is designed in the first instance. Sometimes an exploratory analysis is required because the items do not sufficiently precisely conform to your expected measurement model. The most common causes of this phenomenon are because the model was poorly specified (i.e., it has a poor theoretical underpinning) or because items were poorly constructed (e.g., Booth & Hughes, 2014). However, if it can be used, CFA has many advantages, including: it provides a relatively strong test of the number of factors, it provides relatively sound data on the behavior of items and, according to Joreskög (1969), avoids the problem of factor score indeterminacy. All of these advantages only accrue if the sampling procedures and sample size are adequate, and that sensible choices are made with respect to estimators and fit criteria.

From the perspective of developing scales with good psychometric properties, ideally a CFA model will conform to very simple structure. This is a structure in which items load on just one factor and loadings of items on all other factors are zero. This type of simple structure is different from Thurstone's original conception and for some purposes may not be strictly necessary (Mulaik, 2010). However, if all items provide effectively pure measures of a unidimensional construct this is highly desirable and would be the case if very simple structure holds (Hattie, 1985).

A key decision in CFA is the choice of estimator. Provided its assumptions are met maximum likelihood is a very attractive choice since asymptotically it provides consistent, efficient, and unbiased estimates, that is, estimates with small standard errors, and which tend toward the true population parameters (Bollen, 1989; Larsen & Marx, 1981). However, item-level data is almost never continuous or normally distributed, so in practice for scale development maximum likelihood is almost never the
theoretically correct option. Nevertheless, many authors prefer maximum likelihood to the alternatives arguing that in practice it performs well with binary items and items with at least five response options (Dolan, 1994; Millsap & Kim, Chapter 26). However, at least theoretically, weighted least squares (Browne, 1984) and diagonally weighted least squares are preferable since they are designed for ordered-categorical data and make no assumptions about distribution (Flora & Curran, 2004). In practice, under most circumstances it is diagonally weighted least squares that recovers population parameters best (Flora & Curran, 2004). There is also clear evidence that maximum likelihood does not always perform well with ordered-categorical items (MacCallum, Browne, & Cai, 2012; Nye & Drasgow, 2011). Increasingly, therefore, diagonally weighted least squares is becoming a preferred option.

The issue of model fit is also important if the advantages of CFA are to be realized. There is an extensive literature on this subject much of it focusing on fit when using the maximum likelihood estimator (e.g., Bentler, 1990; Fan & Sivo, 2007; Hu & Bentler, 1999, 1998; Marsh, Hau, & Grayson, 2005; Schermelleh-Engel, Moosbrugger, & Muller, 2003). Classic studies by Hu and Bentler (1998, 1999) that used Monte Carlo simulations have provided empirically derived recommendations with respect both to which fit indices recover the correct solution most consistently and what cut-offs to employ. A somewhat simplified version of their recommendations is that multiple indices of fit should be employed, and that the following fit indices with associated cut-off values perform well; (Root Mean Square Error of Approximation (RMSEA) ≤ 0.06, Standardized Root Mean Square Residual (SRMSR) ≤ 0.08, Tucker–Lewis Index (TLI) ≥ 0.95, Comparative Fit Index (CFI) ≥ 0.95), and application of these golden rules more or less represents current practice. Actually, Hu and Bentler’s recommendations were more nuanced, suggesting that different indices are particularly sensitive to different sources of misspecification. Moreover, many commentators have warned against slavishly following golden rules, and instead suggest that statistical judgment needs to be exercised (e.g., Marsh et al., 2010; Yuan, 2005).

One of the disadvantages of using the diagonally weighted least squares estimator is that there is less simulation data on which to base choices of fit indices and accompanying cut-off values. Nevertheless, Nye and Drasgow (2011) have shown that the golden rules do not work well with diagonally weighted least squares. They recommend that the RMSEA and SRMSR be used and warn that the TLI and CFI perform poorly. Their simulation shows that much more restrictive cut-offs are required, and that these are dependent on a number of factors including sample size. They offer an excel program to calculate the appropriate cut-off values. One limitation is that their simulation only applies to dichotomized data but, clearly, any data can be transformed into this form. Probably a similar approach could be adopted when using the maximum likelihood estimator given that the golden rules often perform poorly in this case too (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Fan & Sivo, 2005, 2007; Yuan, 2005).

It has been recognized for many years that any statistical model is just an approximation and that it would rarely if ever hold exactly (MacCallum, Browne, & Cai, 2012). This is as true of CFA as any other model. However, there is a very general deficiency of CFA that would counsel against its sole use in scale development, which is that a CFA model can show excellent psychometric properties when in fact the scale provides unreliable measurement at some levels of the trait. For this reason alone, there is a strong argument for applying IRT as well as CFA for scale development. IRT has many other advantages, too (Embretson & Reise, 2000, p. 15). One principal advantage is that it
can provide the statistical basis of computer adaptive testing. The latter can provide reliable tests with shorter testing times and is relatively robust to cheating when administered via the internet, without proctoring. In the modern era, these are considerable advantages. For these reasons, most widely used ability and achievement tests (e.g., the SAT, WAIS IV, Woodcock–Johnson IV) are actually based on IRT rather than CFA, although currently, IRT is not commonly used in personality or attitude testing. We will illustrate the uses of both CFA and IRT in our practical example.

Reliability

Psychometric tests are evaluated using three conceptually different estimates of reliability: internal consistency, test-retest, and coefficients of equivalence. Revelle and Condon (Chapter 23) provide an extensive treatment of reliability, here we will make some simplified recommendations. By some distance, the most commonly reported measure of reliability is Cronbach’s alpha. What is not commonly realized is that under most conditions Cronbach’s alpha is a Guttmann lower bound to reliability (Rae, 2007; Zinbarg, Revelle, Yovel, & Li, 2005). Under most circumstances, MacDonald’s Omega provides the most accurate estimate of reliability. Code 2 in the Code Appendix provides example MPlus code to calculate Omega, and Revelle and Condon (Chapter 23) provide R code. Because Cronbach’s alpha is so widely understood we recommend that both it and Omega are routinely reported in technical specifications of tests.

It can be shown that reliability is the ratio of true score variance to total observed variance, or the ratio of true score variance to the sum of true score and error variance. In reality, however, what is regarded as error depends on how a construct is theorized. Possible sources of error variance include the following:

1. Learning, growth, fatigue, forgetting, senescence, biorhythms, maturation, motivation, and historical events
2. Quality of items
3. Extent to which items conform to definition of trait
4. Situational factors: for example, noise, incorrect timing, ambiguous instructions
5. Effects due to (1) practice, (2) rehearsal, and (3) consistency

Schmidt and Hunter (1996) apply a somewhat different trichotomy to sources of error variance:

1. Random response errors that occur within occasions and include variations in attention, mental efficiency, momentary distractions, and so on.
2. Transient errors that vary across occasions and include errors due to mood, feeling, mental efficiency, and so on.
3. Specific errors that consist of variance specific to the tests but are unrelated to the defined trait; for example, item wording, instructions, and items unrelated to trait.

To a degree, internal consistency measures of reliability take account of random response errors and specific errors. However, they clearly do not take account of sources of error that vary over time. For this reason, test-retest or stability coefficients provide useful additional information on reliability. However, these are only appropriate if the
focal construct is hypothesized to show stability over time. It is clearly unsuitable as a measure of the reliability of state measures, which are expected to change over time. Internal consistency measures also do not fully take account of the extent to which items correspond to the trait. For this reason, alternate form or equivalence coefficients are another important source of reliability information. In the case of personality, it is notorious that alternate forms providing measures of purportedly the same construct derived from different personality inventories show particularly poor reliability as measured by equivalence coefficients (Pace & Brannick, 2010). This suggests that different personality measures ostensibly measuring the same trait in fact do so only very approximately. This issue probably affects many other types of measure, although this issue is often not fully explored.

However, for traits used in selection: random response, transient, and specific effects should arguably all be treated as error. Traits used for selection are of little use if they vary much over time, measure irrelevant factors, or are subject to response artifacts. So, Schmidt and Hunter (1996) have argued that the measure of reliability appropriate to selection instruments is the coefficient of equivalence and stability (CES). The CES is estimated by correlating two parallel forms administered at two points in time. While the logic seems compelling, this is yet to become standard practice.

Which coefficients of reliability are actually appropriate to any given test depends on how the construct to be tested is theorized. However, APA, AERA, and NCME (2014) guidelines advise the reporting of internal consistency, stability, and, equivalence measures of reliability, so when developing a test all three should be measured or the technical specification should provide a clear rationale for the inappropriateness of those measures of reliability that are omitted.

Validation

Validation is also an important stage in test development and evaluation, many would argue the key stage (e.g., AERA, APA, NCME, 2014; Smith & Smith, 2005). Hughes (Chapter 24) provides a sophisticated treatment of this issue, while Bollen (1989) and Ghiselli, Campbell and Zedeck (1981) also provide invaluable treatments. Traditional views of validity have evolved over time (Newton & Shaw, 2014) with the current model suggesting that validity is a unitary construct examined using different sources of validity evidence. According to this conceptualization there are not different types of validity. Instead, validity is a summary construct based on varied categories of validity evidence (AERA, APA, NCME, 2014; Messick, 1989). There are broadly five such categories, namely evidence based on test content, response processes, test structure, relationships with other variables, and consequences of testing.

From a pragmatic perspective, it is useful to classify different sources of validity evidence, and as argued by Borsboom, Mellenberg, & van Heerden, (2004), Cizek (2012, 2016), and Hughes (Chapter 24) attempting to understand “validity” as a unitary concept is cumbersome, illogical, and perhaps even impossible. In this Handbook, Hughes (Chapter 24) argues that the types of validity evidence are best considered in response to two questions: Does the test measure what it purports to and is the test useful for some specified purpose. The first of these questions pertains to the accuracy of a psychometric test and the second refers to the appropriateness of a psychometric for a specified purpose. This distinction, though simple, is very useful and also
bypasses one of the major problems often faced by test developers: Whether to maximize either accuracy of measurement or predictive utility. Under a unitary model of validity these two criteria contribute equally toward a single validity coefficient/argument despite the fact that they represent quite different goals that require quite different information and may be mutually exclusive (e.g., Borsboom et al., 2004). Using Hughes’ accuracy and appropriateness model they represent separate goals and processes, so neither needs to be compromised in pursuit of the other (see also Borsboom et al., 2004; Cizek, 2016).

The accuracy of a psychometric test can be established by examining the response processes of the participant while taking the test, the content of the test, and the structure of the test.

Response processes: Psychometric tests are primarily designed to measure constructs (e.g., extraversion, school knowledge) with the underlying assumption that the construct in some way drives the test item response (e.g., those who are extraverted like parties and thus strongly agree with the statement “I like to go to parties”). Thus, if we are trying to measure the construct of extraversion we need to establish that the test item response is driven by extraversion (Borsboom et al., 2004; Embretson, 1984, 1994). There are several methods for investigating item responses: Perhaps the two most notable are cognitive models (e.g., Embretson, 1994, 1998, 2016) and think-aloud protocols (Ericsson & Simon, 1980). In brief, what the test developer needs to do is hypothesize the likely mental processes underlying the item response and then examine whether this is the case. If, in our example, participants strongly agree to the item “I like to go to parties,” but only because they feel this is the appropriate response, then we are measuring social desirability and not extraversion. Thus, our test would be inaccurate.

Test content and structure: One of the most important elements of test construction is ensuring that the test content covers all relevant elements. Given that we are interested in examining test responses (e.g., rotating a shape, retrieving learned information) it is important that test content covers the whole domain of the response process. Thus, if a test is to measure human cognitive abilities then it should assess all known relevant elements of cognitive ability (e.g., Carroll, 1993; Jensen, 1998; Nyborg, 2003). It follows that the stronger the conceptual and empirical base that exists regarding the nature of a construct, the easier it is to apply in the process of test development. If the test systematically samples different and relevant domains of cognitive ability, carefully designs items in accordance with item design guidelines, controls response artifacts and construct irrelevant sources of variance, and makes use of empirical knowledge concerning item content and its relation to known measurement properties, then the test is likely to have content that accurately represents the construct. A related point pertains to the structure of the test content. In simple terms, the test content should match the theoretical model of the construct. If cognitive ability is theorized to consist of a general factor, three third-order factors, 16 second-order factors, and 120 primary factors (Johnson & Bouchard, 2005; McGrew, 2009), then these factors should be identifiable when analyzing test data. If they are identifiable then we can say that the structure accurately matches the theoretical framework posited. As discussed previously, forms of factor analysis provide the most common method for examining the structure of test data.

The second element of validation concerns the appropriateness of using a psychometric test for a given purpose. As with reliability, whether a particular source of validity
Evidence is relevant or not depends on the purpose to which the test score is put and there are many such purposes. In general, whether the use of a test for a specified purpose is appropriate or not is determined by examining the relation of test scores to other variables, the consequences of test use, and the feasibility of test use.

Criterion relations: This is most important in relation to selection tests, where the greatest weight is placed on criterion validity. The rationale is that if a test is to be used to select employees, then the test is useful to the extent that it predicts future performance. For selection tests, therefore, it is usual to measure the correlation between test scores and some measure of job performance, either at the same point in time (concurrent validity evidence) or at some future point in time (predictive validity evidence). The larger the correlation the more appropriate the use of the test for employee selection.

Relations with related and unrelated constructs: Convergent validity evidence refers to the extent to which purported measures of the same construct correlate together, while discriminant validity evidence is focused on the correlations that obtain between different constructs. To the extent that measures of the same trait correlate at one, so may they be said to demonstrate convergent validity evidence. In contrast, it is generally supposed that measures of different constructs should show a lesser degree of correlation. One of the most powerful approaches to the assessment of convergent and divergent validity derives from Campbell and Fiske’s (1959) multitrait-multimethod approach. Koch, Eid, and Lochner, Chapter 25 in volume 2, describe how this initial insight has developed into a class of very sophisticated MTMM models based on SEM. One caution is required: However sophisticated and powerful the MTMM approach may be, it may not suit all types of data. As Koch et al. (Chapter 25) emphasize, which MTMM model is appropriate depends on the constructs of interest and how they are measured. In fact, it is conceivable that for certain purposes MTMM analyses may not be appropriate at all. It is frequently the case that the different methods in MTMM analyses are composed of different raters. Suppose for the sake of argument that the construct of interest is adaptability to different situations, and that each rater sees each participant in only one situation. One index of adaptability would be the extent of change from one situation to the other which would be a difference score. In this situation, the assumption that the extent to which ratings correlate provide convergent validity evidence, which is essentially the principle of MTMM models, would be entirely incorrect. Researchers should consider very carefully what model is appropriate to their data.

Consequences and fairness: One of the key elements of establishing whether test use is appropriate is a consideration of the consequences of test use. In short, tests should not be used if their use is likely to lead to unjustifiable discrimination or unfairness (consequential validity evidence). Both invariance testing and differential item functioning effectively serve the same purpose: To ensure that the test is fair across qualitatively different groups, examples of which might include groups that differ in terms of gender, age, ethnicity, religion, education, or occupation. The purpose of invariance testing is not to show that different groups obtain identical mean scores; there may be real mean score differences across groups (Del Giudice, Booth, & Irwing, 2012); but rather to show that the test scores are not biased. Two chapters in this volume provide a thorough treatment of both methods: Millsap and Kim (Chapter 26 in Volume 2) and Drasgow, Nye, Stark, and Chernyshenko (Chapter 27 in Volume 2).

With regard to invariance testing, it is generally contended that invariance has been demonstrated if the test has the same factor structure across groups (configural
invariance), factor loadings are of identical magnitude (metric invariance), and the items show the same intercept values (scalar invariance: Meredith, 1993; Widaman & Reise, 1997). Appendix Code 3a–3c contains example MPlus code, for the main steps in invariance testing. Some authors also advocate that equality of error variances represents another necessary criterion for invariance to hold (Adolph, Schuurman, Borkenau, Borsboom, & Dolan, 2014). However, Little (2013) provides compelling arguments as to why this is not appropriate.

Differential item testing should, in principle, provide at least highly similar information concerning item and test bias. However, differential item testing is based on IRT modeling. First, a baseline model is estimated and Drasgow et al. (Chapter 27) suggest that the preferable approach is for the baseline model to allow all items to freely vary across groups, with one or more reference items constrained to equivalence. They then suggest that item parameters should be constrained one at a time, and the difference in fit examined. If, in the more constrained model, there is a marked decrement in fit then it is concluded that the constrained item evidences DIF. However, along with Millsap and Kim (Chapter 26), they also suggest that effect size measures of bias are examined. The question is not just whether bias exists but whether it is of practical significance, especially at the test rather than the item level.

Feasibility: One final consideration regarding the appropriateness of test use, not currently contained within standard validation guides (e.g., AERA, NCME, APA, 2014), pertains to the feasibility of test use (Cizek, 2016). This element of validation concerns the rather practical issue of whether or not test use is possible and whether it is sensible according to pragmatic boundaries (e.g., time, cost).

The many forms of validation evidence can be considered using Hughes’ (Chapter 24) accuracy and appropriateness model to help test developers maximize the quality of their tests both in terms of measurement and utility. In sum, if responding to a test item requires the use of the target construct (and thus the item response is derivative of the construct), the test content accurately reflects the full domain of the construct and the structure of the data is as hypothesized then we can say we have an accurate test. If the use of a test for a given purpose is supported by criterion, consequential, and feasibility evidence then we can conclude that the use of a test is appropriate.

Test Scoring and Norming

The choice of which method to use in order to score a test is dealt with in more detail elsewhere in this volume (Brown, Chapter 20; Dorans, Chapter 19). IRT scoring is the norm for large commercial tests of cognitive ability, and it has considerable advantages, not least reduced testing time. However, in many situations this form of scoring is not practical. Here we describe unit weighted scoring schemes, which produce standardized scores using an appropriate standardization sample, variants of which include stanine, sten, and t scores (Smith & Smith, 2005). There are arguments for using either population representative samples or samples that are representative of the applicant pool. In many cases, test developers offer both. For many research scales, simple unit weighted sums of item scores without standardization are considered sufficient.

Scale scores are naturally based on the items that have been previously shown both to measure the focal construct and provide unidimensional, reliable, and unbiased measurement. Raw scale scores can either be based on a unit weighted sum of item scores or
on factor scores. It is generally contended that unit weighting is more generalizable, although that should depend on the size of the standardization sample. Whichever form of scoring is used, the next step is usually to transform the raw scores so that they provide a good approximation to the normal distribution. Recently some sophisticated programs have been developed in order to achieve this, including the R package `fitdist`, which provides four different methods of fitting a variety of the most commonly used distributions to data (Delignette-Muller & Dutang, 2015). However, under most circumstances traditional Box–Cox procedures, which estimate the power to which the scale score should be raised in order to conform to a normal distribution, provide adequate approximations to normality (Box & Cox, 1964). A number of software packages will provide estimates of the power transformations required for scale scores to conform to normality, the “bcskew” algorithm in STATA being an example. Subsequently the scale score should be raised to the appropriate power previously estimated, and then standardized by subtracting the mean transformed score from the transformed scale scores and dividing by the standard deviation of the transformed scores. The resulting standardized scores (z scores) should then be multiplied by the standard deviation and added to the mean appropriate for the type of scale required: For stens (M = 5.5, SD = 2), stanines (M = 5, SD = 2), t scores (M = 50, SD = 10), and intelligence quotients (M = 100, SD = 15). Choice of scoring system is largely down to convention, although stens and stanines are preferred when measurement is more approximate as in the case of most personality scales, while t scores and intelligence quotients are preferred when measurement is highly reliable and a good approximation to continuous measurement (Smith & Smith, 2005).

**Test Specification**

The test specifications can only be prepared once all necessary trialing, validation, and standardization studies have been completed and analyzed, although intermediate specifications will likely be required. At one level, the requirement is fairly simple. You need a list of items and item codes (codes for each item are invaluable, e.g., in databases, when writing syntax, etc.) that comprise each version of the test, using an appropriate program and format for storage. Also required are demographic items and the response formats for each item (e.g., Strongly Agree, etc.). The scoring algorithm should be specified in detail and may take the form of lines of syntax in some high-level computer language. You also need to specify the exact design of the published form of the test, including features such as layout, text formatting, and question formatting, whether it appears in print form, or is administered via the internet.

However, the issue becomes complex because most test developers trial items both within the test and in separate pilot studies. Also, changes can be incremental, so what amounts to a new version of the test, and what is just a minor incremental change requires some consideration. What also complicates this issue is that many modern tests use procedures to preserve test usefulness in the circumstance that test taking is not proctored, and that therefore there is a potential for cheating. Measures to combat this may take the form of computer adaptive testing, or a procedure akin to what Morrison and Embretson (Chapter 3) describe as algorithmic item generation. In the case of computer adaptive testing the required item banks are large, whereas with algorithmic item generation large banks of alternative numerical values for each variable need to be
pre-specified, Although ultimately designing test specifications is usually fairly simple, the most important lesson is that unless systematic attention is devoted to this issue, there can be a considerable cost in terms of confusion and wasted time.

Implementation and Testing

The sophistication of the implementation and testing will depend on the modality of administration and the four dimensions that affect test complexity defined before (e.g., single scale vs. battery, scale similarity, etc.). Whether the test is administered in paper and pencil form or using internet administration packages, some form of checking process is required to ensure that the test has been correctly implemented. In the case of some commercial tests using bespoke software and automatic scoring the testing process may be very thorough and extensive, which leaves many opportunities for test implementation errors to arise.

The MAT^{80}, described next, is a test battery that uses mixed scale types, narrow spectrum measures, and was developed using a small development team. It is implemented as an internet-based test, provides more or less instant scores which are incorporated into an extensive report, using static rather than dynamic text. For the MAT^{80}, the implementation checks were comprehensive and well defined. At a macro level the major objectives of the checks were to ensure that: (1) The test specification was entirely accurate, especially that it listed all items correctly, and that the scoring algorithm was error free; (2) the scoring algorithm had been correctly implemented on the server; (3) that the appearance and functionality of the interface conformed to the specification; (4) that the test worked on a variety of platforms and software; (5) that all reports were correctly generated within 40 seconds; and (6) that all faults were correctly logged and communicated to the appropriate person in order to be rectified. Probably the simplest example is the check on machine scoring of tests. First, a comprehensive list of test items with appropriate codes was compiled. Second, the scoring syntax was implemented in SPSS and repeatedly applied to test data until there were no faults in the scoring process and all resultant scores showed appropriate properties in terms of mean scores, standard deviations, range, kurtosis, and skewness. The repetition was required in order to eliminate errors from hundreds of lines of code. At this point the test specification was developed and checked. Next, the specification was presented to the team responsible for implementation on the server. This team translated the items and the scoring syntax into the appropriate machine language. In order to ensure that the scoring syntax had been correctly implemented on the server, a set of test data was then scored on the server. These scores were compared with scores based on the same data set, but this time scored on the original machine using the scoring syntax taken directly from the test specification. The scoring syntax was deemed to be correctly implemented on the server when both sets of scores were identical.

Subsequent steps involved directly inputting data by responding to all questions generated via the internet test interface. The simplest of these processes involved tapping in standard score patterns and ensuring that the resulting scores were correct. Many other checks were conducted, for example on the interface, and to ensure performance across different platforms (e.g., operating systems and web browsers), that different reports were all correctly produced. While the precise details will vary considerably, for tests to work to an appropriate standard some form of systematic checking procedure is
required. Without a well-defined and systematic checking process, it is unlikely that the test interface or the test scores will be error free.

**Technical Manual**

The technical manual should describe the results of steps 1–9. No commercial test would be saleable without such a manual, and it is also invaluable for all other types of test. An extensive manual is also required for test administrators. The exact form of this manual will vary considerably depending on the nature of the test, but the most important information to be included concerns the interpretation of test scores.

**Example analyses for the Technical Manual**

The example analyses presented here are taken from the MAT$^{80}$, a test designed for recruitment to managerial positions or MBA programs. The full test is comprised of four sections: Section 1 measures Business Personality, Motivation, and Leadership; section 2 assesses Problem Solving, Creativity, and Innovation; section 3 tests Business Numeracy; and section 4 tests Business Reasoning. The subset of analyses presented do not represent a comprehensive account of the analyses conducted in the development of the MAT$^{80}$, but rather are chosen to be illustrative of the type of analyses required in test development.

Prior to the analyses presented here, each scale had been subject to a process of test development including initial construct definition, desk and empirical research designed to elucidate the dimensional structure of the items (Stage 1, Table 1.1), a process of item development and review (Stage 2, Table 1.1), followed by preliminary administration and analysis in two different data sets (Stage 3d, Table 1.1). Without this development, it is unlikely that scales would fit a confirmatory factor model.

The majority of analyses presented were based on a sample of 1,777 applicants to a Global MBA program, which provided the basis of version 1 of the MAT$^{80}$. Selected demographic characteristics of this sample are shown in Table 1.3. However, some of the analyses pertain to version 2 of the MAT$^{80}$. Of the 106 items that make up version 2 of the MAT$^{80}$, 74 items were new. The Business Numeracy items were also refined. This extensive program of item development was undertaken to increase diversity of item content, improve reliability, and minimize the already very low levels of bias.

We will use the accuracy and appropriateness model of test evaluation (Hughes, Chapter 24) to structure the presentation of the analysis. With respect to accuracy we conducted all of the recommended analyses, with the exception of examining response processes. Rather than examining response processes, the accuracy of the tests in representing each focal construct was assessed on the basis of expert review and structured interviews. Concretely based on an extensive knowledge of the literature, for example that contained in Carroll (1993), Booth (2011), the IPIP (Goldberg, 1999), and of course many other sources, we formulated tight construct definitions, and made an expert judgment as to whether scales conformed to these definitions. These evaluations were repeated on numerous occasions, both individually, in groups of two, and in expert panels. For all scales excepting those assessing cognitive ability, we also interviewed approximately 200 members of the public to explore their understanding of each construct, and the language used was incorporated into items. This process is
not identical to the use of thinking aloud protocols, but arguably provides similar and some additional data, in that accuracy was engineered into the items. In addition, the structure of the test content was also examined using factor analysis.

With regard to appropriateness, we primarily investigated discriminant, predictive, and incremental predictive validity evidence as the ability to predict appropriate criteria is the key issue for a selection battery. We also examined the stability of the measure across groups to ensure that test use was fair.

Table 1.4 provides the definition together with an example item for each scale in the Business Personality, Motivation, and Leadership section of the test inventory. The 49 items comprising these eight scales were subject to CFA using the diagonally weighted least squares estimator as implemented in MPlus (Muthén & Muthén, 1998–2010). For example MPlus code, please see Appendix Code 1. The resulting pattern matrix is shown in Table 1.5. The corresponding indices of fit for this model were ($\chi^2 = 6,750.0$, df = 1,099, CFI = 0.905, TLI = 0.899, RMSEA = 0.054). According to conventional cut-off values, the CFI and TLI are indicative of moderate fit, and the RMSEA indicates good fit. It is clear from a variety of simulations that simple rules of thumb with regard to fit do not work (Nye & Drasgow, 2011; Yuan, 2005). This model is also far more complex than CFA models typically used in Monte Carlo

Table 1.3  Demographic characteristics of the norm sample for the MAT$^{80}$.

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Response</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>1,256</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>497</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>24</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>White</td>
<td>326</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>818</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>166</td>
</tr>
<tr>
<td></td>
<td>Arab</td>
<td>268</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>28</td>
</tr>
<tr>
<td>Education</td>
<td>Secondary school to age 16 years</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Secondary school to age 18 years</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Non-university higher education</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>Undergraduate university education</td>
<td>1,094</td>
</tr>
<tr>
<td></td>
<td>Postgraduate university education</td>
<td>534</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>27</td>
</tr>
<tr>
<td>Occupation</td>
<td>Professional/Senior Managerial</td>
<td>1,134</td>
</tr>
<tr>
<td></td>
<td>Junior Professional/Managerial</td>
<td>522</td>
</tr>
<tr>
<td></td>
<td>Administrative/Secretarial</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Skilled</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Semi-unskilled</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Service</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>38</td>
</tr>
<tr>
<td>Age</td>
<td>Mean years</td>
<td>32.7</td>
</tr>
</tbody>
</table>

N = 1,777
simulations (Hu & Bentler, 1998, 1999; Nye & Drasgow, 2011). Strictly then, there is no definitive interpretation of these fit statistics. However, given the complexity of the model, the fit statistics suggest that the level of misspecification is not substantial. The pattern matrix shows that all but three of the 49 loadings are greater than 0.5, 36 are greater than 0.6, and the highest loading is 0.907. That the loadings are substantial and that the pattern matrix matches the theorized structure suggests the items are accurate, that is, 46 of the 49 measure their respective constructs well, and they contain little construct irrelevant variance. The factor correlations range from 0.333 to 0.801 in magnitude. While the constructs are highly correlated, they are nevertheless distinct.

Reliability and scale characteristics For each of the eight scales, a unit weighted sum score was created. In order to reduce skewness and kurtosis, Box-Cox (1964) power transformations were first calculated in STATA and then applied to the sum scores. These transformed scores were then converted into stanines. The scale reliability (McDonald’s Omega and Cronbach’s Alpha), mean, standard deviation, skew, kurtosis, and standard error of measurement for each of the resultant Business, Personality, Motivation, and Leadership scales are presented in Table 1.6. McDonald’s Omega was chosen as the reliability statistic as it generally provides a good estimate of reliability (Revelle & Zinbarg, 2009; Zinbarg, et al., 2005). For example MPlus code, please
Table 1.5  Standardized factor loadings for the Business Personality, Motivation, and Leadership scales.

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.669</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.647</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.658</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.744</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.563</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.547</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.616</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.622</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.559</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.452</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.573</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.660</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>0.448</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>-0.595</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>-0.688</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>-0.731</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td>0.827</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td>-0.651</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td>0.767</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td>0.610</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td>0.783</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td>-0.657</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td>0.699</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td>0.686</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td>0.751</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td>0.375</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td>0.814</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.755</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.722</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.695</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.614</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.907</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.698</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.887</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.790</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.631</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.747</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.525</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.654</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.539</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.622</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.530</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.791</td>
</tr>
</tbody>
</table>

(continued on p. 28)
see Appendix Code 2. Perhaps it is to be expected that the stanine scores are correct, given that a simple numerical formula was applied to obtain them, however, it is also apparent that the Box-Cox transformations provide a good approximation to normality, and even the figures for kurtosis are acceptable. Without using power transformations, it is not possible to achieve all these properties simultaneously. Also provided are the reliabilities for version 2 of these scales (Omega-2), in which as described previously 77% of the items are new, excluding those measuring cognitive ability, which had just been refined. An extensive program of scale development has improved both the diversity of item content and the reliability of the scales, which are impressive.

**Fairness and the Business Personality, Motivation, and Leadership Scales**

The Business Personality, Motivation, and Leadership Scales were tested for fairness across gender and ethnicity using the method of assessing factorial invariance in ordered-categorical measures devised by Millsap and Yun-Tein (2004). All analyses were conducted using the diagonally weighted least squares estimator as implemented in MPlus (Muthén & Muthén, 2010). Testing for invariance provided a simultaneous test of the suitability of the MAT for candidates for whom English is a second language. For example MPlus code, please see Appendix Code 3a–3c.

Various criteria have been suggested for the measurement of fairness. In a recent simulation study Chen (2007) suggested a two-step criterion, which differs depending on the magnitude of the sample and the evenness of sample sizes. For sample sizes of 300 or
less and uneven sample sizes Chen suggests that fairness requires that the CFI does not reduce by .005 or more and the RMSEA does not increase by more than .01 when: (1) comparing the metrically invariant to the configurally invariant model and (2) the scalar to the metrically invariant model. When sample sizes are 500 or more and are even in size then fairness requires that the CFI does not reduce by .010 or more and the RMSEA does not increase by more than .015.

The results of the invariance testing for ethnicity and gender for the Business Personality, Motivation, and Leadership scales are presented in Table 1.7. When considering the Business Personality, Motivation, and Leadership scales, the most stringent criterion only applies to the comparison of Whites with South Asians. Here the CFI increased by .001 (Metric vs. Configural), and then decreased by .001 (Scalar vs. Metric), and the RMSEA increased by .001 for both comparisons (see Table 1.7). These results are substantially within the stringent criteria specified by Chen and show convincingly that the Business Personality, Motivation, and Leadership scales are fair when comparing South Asians with Whites. Similarly, for the remaining comparisons across ethnicity and gender, the less stringent criteria for fairness were met in every case (see Table 1.7). In fact, considered as a whole in 6 out of 10 cases the CFI either increased in magnitude or showed no change and the RMSEA decreased; that is, the metrically and scalar invariant models showed a better fit than the configurally invariant model. Improvement in fit of more restrictive models is very unusual, and provides convincing evidence of the fairness of the Business Personality, Motivation, and Leadership Scales.

The MAT80 was devised using an explicit language standard, including preferred use of the 850 most common English words wherever possible, explicit rules concerning the
simplification of language, as well as highly explicit item writing guidelines. This approach to item development may explain why our test items do not appear to show bias across different ethnic groups residing in different countries.

Predictive validity evidence

In order to establish the predictive validity of the MAT$^{80}$, a number of analyses were undertaken. First, a series of SEM were estimated in order to determine the MAT$^{80}$'s ability to predict overall semester marks for MBA students. Second, these multiple Rs were combined using meta-analytic techniques to provide an overall estimate of the MAT$^{80}$'s predictive validity.

This approach to examining predictive validity differs from standard practice in two major respects. Firstly, the most common form of educational outcome data used to assess predictive validity of the Graduate Management Admission Test (GMAT), the most widely used selection test for MBA programs, is first-year grade point average, with graduate grade point average also used, but much less commonly. The disadvantage of both these outcome criteria is that the factor structure of these data is assumed rather than measured and there is no correction for unreliability of measurement. These difficulties with respect to standard practice reflect a subset of a complex range of problems with outcome criteria that have been collectively labeled the criterion problem (e.g., Austin & Villanova, 1992; Viswesvaran & Ones, 2000).

When we applied structural equation modeling to our outcome data we found that not all measures were reliable indicators, and that not all indicators loaded on a general factor. So, it may be concluded that the SEM models more accurately represent the criterion data. A disadvantage of this approach is that, because students take different courses, we cannot take a simple sum of grade scores in order to obtain an overall grade point average. Instead, we combined the estimates of multiple Rs using a weighted average, as is standard in psychometric meta-analysis (Hunter & Schmidt, 2004). However, here we are averaging across different outcomes and different subsets of students. While the overall point estimate is probably accurate, it is based on a missing data design with unknown missingness characteristics. Table 1.8 presents the results of this analysis.

<table>
<thead>
<tr>
<th>Semester marks</th>
<th>N</th>
<th>RMSEA</th>
<th>SRMR</th>
<th>$\rho$</th>
<th>Prange res.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>734</td>
<td>.028</td>
<td>.024</td>
<td>.283</td>
<td>.355</td>
</tr>
<tr>
<td>2nd</td>
<td>608</td>
<td>.042</td>
<td>.026</td>
<td>.390</td>
<td>.470</td>
</tr>
<tr>
<td>3rd</td>
<td>368</td>
<td>.051</td>
<td>.038</td>
<td>.494</td>
<td>.630</td>
</tr>
<tr>
<td>4th</td>
<td>355</td>
<td>.040</td>
<td>.037</td>
<td>.377</td>
<td>.503</td>
</tr>
<tr>
<td>5th</td>
<td>294</td>
<td>.034</td>
<td>.021</td>
<td>.421</td>
<td>.553</td>
</tr>
<tr>
<td>All</td>
<td>1,599</td>
<td>NA</td>
<td>NA</td>
<td>.393</td>
<td>.507</td>
</tr>
<tr>
<td>Semesters 2–5</td>
<td>1,599</td>
<td>NA</td>
<td>NA</td>
<td>.420</td>
<td>.552</td>
</tr>
</tbody>
</table>

Note: Course marks for the semester are based on the courses taken in each semester. For Semesters 1–5 the courses are: 1st Semester: Marketing, Managerial Economics, Global Events and Leadership; 2nd Semester: Accounting in Business, Operations Management; 3rd Semester: Comparative and International Business, Corporate Finance, Business Simulation; 4th Semester: Strategic Management, Final Project Preparation; 5th Semester: People Management and Organisations. RMSEA = Root mean square residual; SRMR = SRMSR; $\rho$ = multiple R corrected for unreliability in the criterion, Prange res. = Multiple R corrected for range restriction.
The RMSEA and SRMR estimates demonstrate that all models estimated were a close fit to the data (see Table 1.8) (for the RMSEA ≤ 0.06 = close fit, > 0.06–0.08 = good fit, 0.08–0.10 = acceptable fit; for the SRMR ≤ 0.05 = close fit, > 0.05–0.08 = good fit; Hu & Bentler, 1998, 1999; Schermelleh-Engel, Moosbrugger, & Muller, 2003).

Based on extensive experience of university examining, we suggest that Semester 1 marks are rarely a reliable indicator of future performance since the majority of students at this stage are going through a period of adjustment to the requirements of higher education. Examining the predictive validities in Table 1.8 also suggests that Semester 1 marks represent an outlier. Therefore, arguably the best estimate of the overall predictive validity of the MAT80 is based on the mean for Semesters 2–5. As can be seen in Table 1.8, the results of this analysis (e.g., MPlus code, please see Appendix Code 4) suggest that the overall predictive validity of the MAT80 in predicting MBA students’ marks in Semesters 2–5 is 0.42 once unreliability in the criterion has been accounted for.

It is of note that the highest estimated predictive validity for the MAT80, at 0.63, occurred when students undertook courses in Comparative and International Business, Corporate Finance, and Business Simulation (Semester 3). It appears that the MAT80 is most capable of predicting performance in those subjects most closely related to core business skills.

The predictive validity of the MAT80 can be understood further by considering it against the predictive validity of the GMAT, currently the most popular psychometric test utilized by business schools for admissions to graduate management programs. Table 1.9 presents comparison data for the GMAT taken from Kuncel, Credé, and Thomas’ (2007) meta-analysis.

In order to determine the most appropriate grounds for comparison of the MAT80 and the GMAT, it is first necessary to consider that the MAT80 is primarily administered to applicants for whom English is a second language. This means that the most closely equivalent data for the GMAT is that for applicants who are non-native English speakers. The estimate of the overall predictive validity of the MAT80, at 0.42, thus

Table 1.9 Meta-analysis of GMAT validities with corrections for unreliability in the criterion and range restriction.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>N</th>
<th>Corrected for unreliability</th>
<th>Corrected for unreliability and range restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Native English-speaking</td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>48,915</td>
<td>.253</td>
<td>.340</td>
</tr>
<tr>
<td>Quantitative</td>
<td>48,758</td>
<td>.325</td>
<td>.380</td>
</tr>
<tr>
<td>Total</td>
<td>28,624</td>
<td>.344</td>
<td>.470</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-native English-speaking</td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>1,815</td>
<td>.157</td>
<td>.210</td>
</tr>
<tr>
<td>Quantitative</td>
<td>1,815</td>
<td>.299</td>
<td>.350</td>
</tr>
<tr>
<td>Total</td>
<td>1,815</td>
<td>.300*</td>
<td>.374*</td>
</tr>
</tbody>
</table>

* Estimate.

Note. Adapted from “A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance.” By Kuncel, N. R., Credé, M., Thomas, L. L. Academy of Management Learning & Education, 6, 60. Copyright 2007 by Academy of Management. Adapted with permission.
compares favorably to the equivalent GMAT estimate, which stands at 0.30. Importantly, however, the MAT\textsuperscript{80} also appears to enjoy a sizable advantage over the GMAT with respect to predictive validity in English-speaking samples.

It would be useful to know the true predictive validity of the MAT\textsuperscript{80} corrected for range restriction, and to compare that with the equivalent figure for the GMAT. One conservative approach is to use the same correction for direct range restriction as Table 1.9.

Applied by Kuncel et al. (2007). The correction for direct range restriction uses the ratio of the standard deviation (SD) from the unrestricted sample to that in the restricted sample. In the case of Kuncel et al. (2007), the unrestricted sample was the total applicant population to all business schools using the GMAT and the restricted sample was successful candidates to all business schools in the sample. Arguably, such a correction provides a lower bound correction for the MAT\textsuperscript{80}. Given the highly selective nature of the MBA program studied, it is likely that the successful applicants represent a more selected sample than is true for the full range of business schools. Thus, it is reasonable to assume that in such a sample the effects of range restriction would likely be greater than is true for the total GMAT sample, which includes the total range of business schools from top to bottom. If one applies corrections in this way then the best estimate of the mean predictive validity of the MAT\textsuperscript{80} is 0.55, whereas for the GMAT the estimate is 0.37 when administered to non-native English speakers, and 0.47 when administered to native English speakers.

The results of the predictive validity analyses thus suggest the following two conclusions. Firstly, on any reasonable comparison the MAT\textsuperscript{80} appears to have a higher predictive validity than the GMAT. Secondly, even at a conservative estimate of 0.55, the predictive validity of the MAT\textsuperscript{80} is strong by conventional criteria (Cohen, 1988). In 2008, Oh, Schmidt, Schaffer, and Lee applied a new indirect range restriction correction to the GMAT data of Kuncel et al. (2007) and concluded that the validities had been underestimated by 7%. Making this correction, the overall predictive validity of the GMAT in English-speaking samples is 0.50, and for the MAT\textsuperscript{80} the revised figure is 0.59, but for predominantly non-English-speaking samples, in which the GMAT performs poorly.

Illustrative IRT analysis of the Leadership scale

The MAT\textsuperscript{80} was not developed using IRT, with the sole exception of the Business Numeracy scale. This broadly follows convention in that until relatively recently IRT has not routinely been applied to the development of personality-type items. However, as argued previously, IRT has many advantages over classical test theory (see Brown, Chapter 21; Embretson & Reise, 2000).

Here we provide an illustrative analysis of the current measure of Strategic Leadership used in the MAT\textsuperscript{80}. Four of ten items are shown in Table 1.10. Notably the items are diverse in content with no repetition, something lacking in many scales of this type. Here we used responses from 2,360 applicants to a Global MBA program. The demographic characteristics of the sample are very similar to those shown in Table 1.3.

The fit of this data to a unidimensional factor model was reasonable, although we would have preferred a smaller value for the RMSEA ($\chi^2 = 813.2$, df = 35, CFI = 0.951, TLI = 0.937, RMSEA = 0.097). We therefore, estimated the item parameters
according to Samejima’s Graded Response model, assuming unidimensionality, with
the program Multilog version 7.0.1, using the Marginal Maximum Likelihood estima-
tor. Figure 1.2 shows the item characteristic curves. It is apparent that the symmetrical
pattern desired for these curves is not attained (see Chapters 15 and 16 for further IRT
discussion). This is because the responses show a marked negative skew. Figure 1.3
shows the test information function and standard errors of measurement over the range
\( z = -3 \text{ to } +3 \). Using this information and applying formula 20.19 from Brown
(Chapter 20), the reliability of measurement at each trait level in the range \( z = -3 \text{ to } -1.6 \), varies from a low of 0.93 to a high of 0.97. If candidates, in order to be selected,
had to attain a z-score of 1.6, that would represent a selection ratio of 1:18.2. It would
be unusual in practical selection contexts for the selection ratio to exceed 18.2. As
Schmidt and Hunter (1998, p. 263) observe, “actual selection ratios are typically in
the 0.30–0.70 range.” In consequence, the reliability of the Strategic Leadership scale

Table 1.10  Example items from the Strategic Leadership scale.

<table>
<thead>
<tr>
<th>Item</th>
<th>1–4</th>
<th>5–8</th>
<th>9–10</th>
</tr>
</thead>
<tbody>
<tr>
<td>People often look to me to make a decision</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I often see future developments before others</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>People always do their best work for me</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I really understand people’s strengths</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.2  Item characteristic curves for the 10 items of the Strategic Leadership scale under the
graded response model.
would be regarded as excellent for most feasible selection decisions, according to conventional criteria. Although theoretically had we strengthened the wording of items we might have obtained a less skewed response distribution, in practice, skewing of responses in high stakes testing is universally obtained when using response options ranging from Strongly Agree to Strong Disagree as is the case here. So, we may conclude that although theoretically IRT should have advantages, in this specific instance little would have been gained by using IRT. For IRT to show an advantage we would have had to use a different testing strategy.

The major omissions from the analyses presented here, which should be included in any scale development, are tests of convergent validity together with tests of test-retest and equivalence reliability. Unfortunately, we are still in the process of gathering data in order to conduct such tests with the MAT.

References


Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F.


Code Appendix

Code 1   MPlus eight-factor CFA model using the WLSMV estimator.

TITLE: CFA of Business Personality, Motivation and Leadership scales

DATA:
FILE IS "D:\Rdata\MAT-80.dat";

VARIABLE:
NAMES ARE act1-act6 ass1-ass4 cuyl-cuy7 fly1-fly6
iln1-iln6 img1-img6 imm1-imm6 incl-inc6 lep1-lep7
opm1 opm2 orr1-orr6 ory1-ory7 prg1-prg6 shg1-shg5
stcl-stc3 sts1-sts6;
USEVARIABLES ARE act1-act6 ass1-ass4 cuyl-cuy7
incl-inc6 lep1-lep7 stcl-stc3 orr1-orr6 sts1-sts6
fly2 fly3 fly5 fly6 prg1 img4 img6
iln6 opm1 opm2;
CATEGORICAL ARE ALL;
MISSING ARE ALL (-999);

MODEL:
ACT BY act1-act6;
ASS BY ass1-ass4;
CUY BY cuyl-cuy7;
INC BY incl-inc6;
LEAD BY lep1-lep7 stcl-stc3;
ORDER BY orr1-orr6;
STRESSRS BY sts1-sts6;
OPTIMISM BY fly2 fly3 fly5 fly6 prg1 act1 img4 img6
iln6 stcl opm1 opm2;

ANALYSIS:
TYPE IS GENERAL;
ESTIMATOR IS WLSMV;
PARAMETERIZATION = THETA;
ITERATIONS = 1000;
CONVERGENCE = 0.00005;

OUTPUT: SAMPSTAT MODINDICES STANDARDIZED (STDYX) RESIDUAL;
**Code 2  McDonald’s Omega for Optimism scale.**

**TITLE:**  McDonald’s Omega for Optimism

**DATA:**

FILE IS "D:\Rdata\MAT-80-3.dat";
FORMAT IS 89F5.2;

**VARIABLE:**

NAMES ARE act1-act6 ass1-ass4 cuyl-cuy7 flyl-fly6
  iln1-iln6 img1-img6 imn1-imn6 inc1-inc6 lep1-lep7
  opm1 opm2 orr1-orr6 oryl-ory7 prgl-prg6 shg1-shg5
  stcl-stc3 stsl-sts6;

USEVARIABLES ARE fly2 fly3 fly5 fly6 prg1 act1 img4 img6
  iln6 stcl opm1 opm2;
MISSING ARE ALL (-999);

**ANALYSIS:**

TYPE IS GENERAL;
ESTIMATOR IS ML;
BOOTSTRAP = 2000;

**MODEL:**

f1 BY fly2*(p1)
  fly3(p2)
  fly5(p3)
  fly6(p4)
  prg1(p5)
  act1 (p6)
  img4 (p7)
  img6(p8)
  iln6(p9)
  stcl(p10)
  opm1 (p11)
  opm2(p12);
f1@1;
fly2(r1)
fly3(r2)
fly5(r3)
fly6(r4)
prg1(r5)
act1 (r6)
img4 (r7)
img6(r8)
iln6(r9)
stcl(r10)
opm1 (r11)
MODEL CONSTRAINT:

NEW(omega);
omega=
(p1+p2+p3+p4+p5+p6+p7+p8+p9+p10+p11+p12)^2/
((p1+p2+p3+p4+p5+p6+p7+p8+p9+p10+p11+p12)^2+
(r1+r2+r3+r4+r5+r6+r7+r8+r9+r10+r11+r12));

OUTPUT: SAMPSTAT STANDARDIZED (STDYX) RESIDUAL;

---

**Code 3** MPlus invariance tests for Business Personality, Motivation, and Leadership scales across gender.

(a) Configural invariance.

**TITLE:** Test of Business Personality, Motivation and Leadership scales for configural invariance across Gender.

**DATA:**

FILE IS "D:\Rdata\mat-80-gender.dat";
FORMAT IS 90F5.2;

**VARIABLE:**

NAMES ARE act1-act6 ass1-ass4 cuy1-cuy7 fly1-fly6 iln1-iln6 img1-img6 imm1-imm6 inc1-inc6 lep1-lep7 opm1 opm2 orr1-orr6 oryl-oryy7 prgl-prg6 shgl-shg5 stcl-stc3 stsl-sts6 gender;
USEVARIABLES ARE act1-act6 ass1-ass4 cuy1-cuy5 cuy7 inc1-inc6 lep1-lep7 stcl-stc3 orr2-orr5 stsl-sts6 fly2 fly3 fly5 fly6 prgl img4 img6 iln6 opm1 opm2;
CATEGORICAL ARE ALL;
MISSING ARE ALL (-999);
GROUPING IS gender(1 = male 2 = female);

**MODEL:**

ACT BY act1-act6;
ASS BY ass1-ass4;
CUY BY cuy1-cuy7;
INC BY inc1-inc6;
LEAD BY lep1-lep7 stc1-stc3;
ORDER By orr1-orr6;
STRESSRS BY sts1-sts6;
OPTIMISM By fly2 fly3 fly5 fly6 prg1 act1 img4 img6
iln6 stc1 opm1 opm2;

MODEL female:

ACT BY act1-act6;
ASS BY ass1-ass4;
CUY BY cuy1-cuy7;
INC BY inc1-inc6;
LEAD BY lep1-lep7 stc1-stc3;
ORDER By orr1-orr6;
STRESSRS BY sts1-sts6;
OPTIMISM By fly2 fly3 fly5 fly6 prg1 act1 img4 img6
iln6 stc1 opm1 opm2;

[lep1$3 lep2$2 lep2$3];
[lep3$2 lep3$3 lep4$2 lep4$3];
[lep5$2 lep5$3 lep6$2 lep6$3];
[lep7$2 lep7$3 stc1$2 stc1$3];
[stc3$2 stc3$3];
[ass1$3 ass2$2 ass2$3 ass3$2 ass3$3 ass4$2 ass4$3];
[sts1$3 sts2$2 sts2$3 sts3$2 sts3$3];
[sts4$2 sts4$3 sts5$2 sts5$3];
[sts6$2 sts6$3 act1$3];
[act2$2 act2$3 act3$2 act3$3 act4$2 act4$3 act5$2 act5$3];
[act6$2 act6$3 opm1$2 opm1$3 opm2$2 opm2$3];
[fly2$3 fly3$2 fly3$3];
[fly5$2 fly5$3 fly6$2 fly6$3 prg1$2 prg1$3 img4$2 img4$3];
[img6$2 img6$3 iln6$2 iln6$3];
[orr2$3 orr3$2 orr3$3];
[orr4$2 orr4$3 orr5$2 orr5$3];
[cuy1$3 cuy2$2 cuy2$3 cuy3$2 cuy3$3 cuy4$2 cuy4$3];
[cuy5$2 cuy5$3];
[cuy7$2 cuy7$3 inc1$3 inc2$2 inc2$3 inc3$2 inc3$3];
[inc4$2 inc4$3 inc5$2 inc5$3 inc6$2 inc6$3];

ANALYSIS:
TYPE IS GENERAL;
ESTIMATOR IS WLSMV;
PARAMETERIZATION = THETA;
ITERATIONS = 1000;
CONVERGENCE = 0.00005;

OUTPUT: SAMPSTAT MODINDICES STANDARDIZED (STDDYX)
RESIDUAL;

(b) Metric invariance.
TITLE: Test of Business Personality, Motivation and
Leadership scales for configural invariance across
Gender.

DATA:
FILE IS "D:\Rdata\mat-80-gender.dat";
FORMAT IS 90F5.2;

VARIABLE:
  NAMES ARE act1-act6 ass1-ass4 cuyl-cuy7 fly1-fly6
  iln1-iln6 img1-img6 imm1-imm6 inc1-inc6 lep1-lep7
  opm1 opm2 orr1-orr6 oryl-ory7 prgl-prg6 shg1-shg5
  stcl-stc3 stts-sts6 gender;
  USEVARIABLES ARE act1-act6 ass1-ass4 cuyl-cuy5 cuyl7
  inc1-inc6 lep1-lep7 stcl-stc3 orr2-orr5 stts-sts6
  fly2 fly3 fly5 fly6 prgl img4 img6
  iln6 opm1 opm2;
  CATEGORICAL ARE ALL;
  MISSING ARE ALL (-999);
  GROUPING IS gender(1 = male 2 = female);

MODEL:

  ACT BY act1-act6;
  ASS BY ass1-ass4;
  CUY BY cuyl-cuy7;
  INC BY inc1-inc6;
  LEAD BY lep1-lep7 stcl-stc3;
  ORDER BY orr1-orr6;
  STRESSRS BY stts-sts6;
  OPTIMISM By fly2 fly3 fly5 fly6 prgl act1 img4 img6
  iln6 stcl opm1 opm2;

MODEL female:

  [lep1$3 lep2$2 lep2$3];
  [lep3$2 lep3$3 lep4$2 lep4$3];
  [lep5$2 lep5$3 lep6$2 lep6$3];
  [lep7$2 lep7$3 stcl$2 stcl$3];
ANALYSIS:
TYPE IS GENERAL;
ESTIMATOR IS WLSMV;
PARAMETERIZATION = THETA;
ITERATIONS = 1000;
CONVERGENCE = 0.00005;
OUTPUT: SAMPSTAT MODINDICES STANDARDIZED (STDDYX) RESIDUAL;

(c) Scalar invariance.
TITLE: Test of Business Personality, Motivation and Leadership scales for configural invariance across Gender.

DATA:
FILE IS "D:\Rdata\mat-80-gender.dat";
FORMAT IS 90F5.2;

VARIABLE:
NAMES ARE act1-act6 ass1-ass4 cuy1-cuy7 fly1-fly6 iln1-iln6 img1-img6 imm1-imm6 inc1-inc6 lep1-lep7 opm1 opm2 orr1-orr6 ory1-ory7 prg1-prg6 shg1-shg5 stc1-stc3 sts1-sts6 gender;
USEVARIABLES ARE act1-act6 ass1-ass4 cuy1-cuy5 cuy7 inc1-inc6 lep1-lep7 stc1-stc3 orr2-orr5 sts1-sts6 fly2 fly3 fly5 fly6 prg1 img4 img6;
iln6 opm1 opm2;
CATEGORICAL ARE ALL;
MISSING ARE ALL (-999);
GROUPING IS gender(1 = male 2 = female);

MODEL:
ACT BY act1-act6;
ASS BY ass1-ass4;
CUY BY cuy1-cuy7;
INC BY inc1-inc6;
LEAD BY lep1-lep7 stc1-stc3;
ORDER By orr1-orr6;
STRESSRS BY sts1-sts6;
OPTIMISM By fly2 fly3 fly5 fly6 prg1 act1 img4 img6
iln6 stc1 opm1 opm2;
MODEL female:

ANALYSIS:
TYPE IS GENERAL;
ESTIMATOR IS WLSMV;
PARAMETERIZATION = THETA;
ITERATIONS = 1000;
CONVERGENCE = 0.00005;

OUTPUT: SAMPSTAT MODINDICES STANDARDIZED (STDXY)
RESIDUAL;

---

**Code 4**  MPlus prediction of 3rd Semester grades using Mat	extsuperscript{80} scores.

**TITLE:** Regression of semester grades on Mat	extsuperscript{80} scores

**DATA:**
FILE IS "D:\Rdata\mat10.DAT";
FORMAT IS 52(F8.4);
LISTWISE=ON;

**VARIABLE:**
NAMES ARE lead opt order assert stress ach cur intrinsc fluency
orig ill prod shar implmnt math markass markwork markex
aiass aibwork aibexam cfass cfwork cfexam cibass
cibwork cibexam gelass gelwork meass mework meexam
nssass nsswork omanass omanwork omanexam pmass pmass2
pmexam pmoass pnowork pmoexam vcpeass vcpeework vcpework smanass
smanwork smanexam bsimass bsimass2 ass1fpp ass2fpp;
USEVARIABLES ARE cfass cfexam cibass cibexam opt assert intrinsc math;
MISSING = ALL (101);

MODEL:
  F1 BY cfass cfexam cibass cibexam;
  F1 ON opt assert intrinsc math;

ANALYSIS:
  TYPE IS GENERAL;
  ESTIMATOR IS MLR;
  ITERATIONS = 1000;
  CONVERGENCE = 0.00005;

OUTPUT: SAMPSTAT STANDARDIZED (STDY) MOD(ALL 5);