Chapter 11
A Problematic Leap in the Use of Test Data: From Performance to Inference

Gabrielle Matters

Introduction

Despite all the rhetoric about the new millennium, few assessment issues thus far belong exclusively to the 21st century. An issue spilling over from the 20th century is the demand for schools and teachers to use assessment information to improve student achievement and enhance educational systems more generally. Among the myriad possible mechanisms for improving student achievement through the efficient use of assessment information by schools and teachers is feedback to the student learning process along with enhancement of teachers' pedagogical repertoires. When the assessment instrument is a standardised test, the product (student responses) gives information not only about what was learnt and how well it was learnt but also about what was not learnt and hints as to why this might be so.

The first section in this chapter provides an organisational framework for description of the generation of assessment data, and applies that framework to standardised testing, focusing on the interactions between student (and student dimensions) and tests (and test items). The section includes a typology for classifying sources of item difficulty. The second section discusses the efficient use of assessment information. It promotes the view that the use of test data by time-poor but intellectually and professionally curious teachers, while requiring rigour, can be a creative and imaginative process. The third section challenges the prevailing way of operating in a world that is 'awash with data' (Hattie, 2005, p. 11), but uncritical of test construct.

The concept of test *construct*, not to be confused with the act of test *construction*, is 'a psychological characteristic (e.g., numerical ability, spatial ability, introversion and anxiety) considered to vary across individuals. A construct (sometimes called a latent variable) is not directly observable; rather, it is a theoretical concept derived from research and other experience that has been constructed to explain observable patterns. When test scores are interpreted by using a construct, the scores

G. Matters (✉)
Australian Council for Educational Research, Suite 1, 165 Kelvin Grove Road, Kelvin Grove QLD 4059, Australia
e-mail: matters@acer.edu.au

C. Wyatt-Smith, J.J. Cumming (eds.), *Educational Assessment in the 21st Century*,
DOI 10.1007/978-1-4020-9964-9_11, © Springer Science+Business Media B.V. 2009

are placed in a conceptual framework' (American Education Research Association, American Psychological Association, & National Council on Measurement of Education, 1985, p. 90).

Ultimately, assessment involves making inferences about student achievement on the basis of the evidence available. One of the essential leaps in the assessment process is from performance to inference (that is, scoring the underlying attribute from what students do). Theoretically, this leap is problematic but most approaches to the use of test data fail to problematise it. Accordingly, in this chapter, I point teachers and schools to talking about the significance of student responses at the item level and seeing what it is that each item actually measures before necessarily concluding from the evidence of a low score on, say, a mathematics test (just a score derived from a collection of items), that the student actually knows no mathematics.

Methodologically, to approach test results at this level could be helpful to teachers and schools because it is not at the level of abstraction of ability: it is about what teachers have to do with their students; that is, to identify the things that students can do, the things they cannot do and things they have trouble with, and understand the source of difficulty.

Reference is made to how this approach could be used in specific forms of external standardised tests. Special mention is made of the Programme for International Student Assessment (PISA) because of the significance that PISA has gained in countries in all corners of the globe (see <http://www.pisa.oecd.org>). By design, PISA assesses the 'aptitude to undertake tasks found in everyday life' (OECD, 2001, p. 20).

The Generation of Assessment Information

This section sets the scene for discussion of schools' and teachers' use of assessment information, with a framework for describing the generation of assessment data and for interpreting patterns and relationships in data. The model is then applied to a specific assessment situation, standardised testing. The section focuses on the student–item interaction through a discussion of student characteristics and features of the testing process that might affect test results.

Organisational Framework in an Assessment Situation

The framework is any adaptation of the 3P model of learning and teaching (Biggs, 1999; Biggs & Moore, 1993), which portrays learning as an interactive system, identifying 'three points at which learning-related factors are placed: *presage*, before learning takes place; *process*, during learning; and *product*, the outcome of learning' (Biggs, 1999, p. 18).

The linear progression from presage to process to product tracks the characteristics of the student that exist before the student enters the learning situation

(plus environmental factors related to the institution, the teacher and the curriculum) through the student's engagement with the learning environment to the outcome —'how much was learned, how well and in what way' (Biggs, 1993, p. 76).

Although teachers and schools have a causally central role in the learning process, students are equally causally central. Student as serious variable in student's own learning is not just student as self. 'Each student is an amalgam of their genetic code and everything that has influenced them. And they continue to be shaped by current influences, both internal and external to the school' (Ericson & Ellett, 2002). In the spotlight in this section is the individual student in the assessment process.

The presage–process–product model, which could also be viewed as a before–during–after model, is adapted to create a framework for describing the generation of assessment data (see Fig. 11.1). In a given assessment situation, for a given student and a given assessment instrument, say a standardised test, there is a definite product—the student's responses, which can be measured (test score); that is, information on what was learnt and how well or, what was not learnt with hints as to why.

These stages tend to be multifaceted, so I focus on one or two particular facets in each stage. For *presage* to the assessment experience (here, testing), I take student characteristics; for assessment *process*, the interaction between the student and item on the test and, for the assessment *product*, test responses and test scores.

This model differs from that used to describe the classroom learning situation (there are different labels on the components for a start). However, it remains a useful model, one that is capable of generating predictions and of providing feedback, both of which are relevant to the study of assessment information.

Elements in bold typeface in Fig. 11.1 are further elaborated later in this chapter. An understanding of them, or indeed, of the elements not in bold type is not necessary at this stage.

Application of Organisational Framework to Standardised Testing

Standardised testing is taken to be the process of administering a test that is the same for all students in the testing population (for example, from group of countries to group of schools to group of students in a subject) or a wide cross-section thereof, taken under the same conditions and marked according to a commonly applied rubric (such as the key for multiple-choice questions, or marking scheme for constructed-response items).

In this section, I choose the external standardised test as the specific form of assessment instrument to illustrate the application of the general organisational framework. The reason for this choice is the significance of PISA in many countries. A quick glance through the presage elements in Fig. 11.1 reminds us of many of the explanatory factors that have come up in conversations and articles about high- and low-scoring countries.

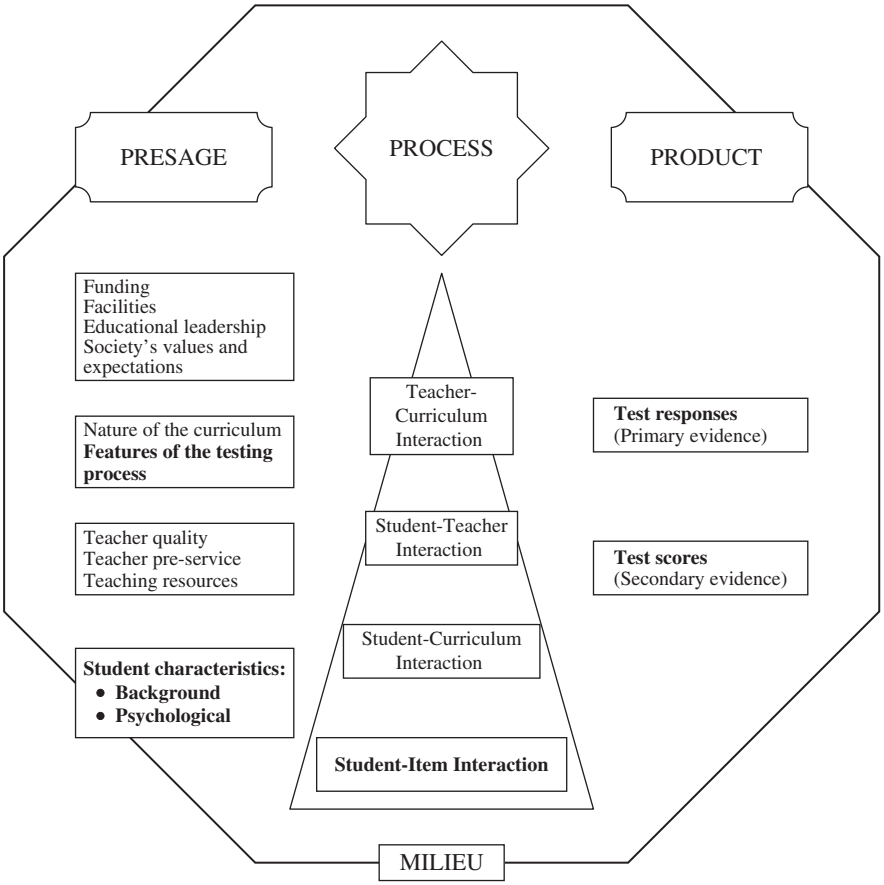

Fig. 11.1 Organisational framework in a testing situation

From Presage to Product

This section focuses on the interaction of student with test item. The student–item interaction is in the 'process' component of the 3P model.

One of the most creative uses of assessment information looks at the products of the assessment process (output data that relate to student achievement) and takes note of how they link with the presage component (input data such as student characteristics and features of the testing process). Examples include the effect of psychological characteristics on test-taking behaviour and therefore on success on tests, the effect of teacher quality on test scores, and gender differences in achievement on different test formats. Student characteristics (background and psychological) are now discussed.

Dimensions of the Student

Students come from different backgrounds, are of different abilities, go to different schools (and keep company with different kinds of children) and have different levels of test preparedness. They also experience different kinds of test items ('hard' compared with 'easy'; 'open' compared with 'closed'), as well as differences in their sources and levels of extrinsic and intrinsic motivation.

Of the myriad characteristics that define an individual at a certain point in that person's life, there are some that the person is born with and others that are the product of a person's environment, in this case, home plus school. Whatever the relative contributions of nature and nurture to the formula for ability, ability is not the same thing as achievement, nor is it the same thing as aptitude, even though doing well at school (academic achievement) is a function of ability, and those who do well at school are taken to have an aptitude for education at this level and the next level. In a nutshell, achievement is what you did, ability is what you could have done and aptitude is what you might be able to do. Ability is the first of the student background characteristics of interest when looking for explanations of patterns, trends and relationships in data.

Other background characteristics that are often included in datasets about students are gender, type of school attended and ethnicity. Although the list is not intended to be exhaustive, the exclusion of socio-economic status (SES) is a deliberate decision. It is my opinion that, on an ethical level, we should refuse to include, at the outset, SES as belonging to the presage component. Otherwise the notion of causality would lead us to the inevitability of low achievement from students from low SES backgrounds.

Achievement is influenced by factors internal to the student as well as to those imposed by features of the assessment environment, which include the assessment instrument itself, preparation for it and conditions under which it is applied. One cluster of internal factors includes the psychological characteristics of the student. Each of the psychological characteristics appearing in the following list is likely to have an impact on the student–item interaction, and therefore the potential to influence the outcome in terms of the quality or accuracy of the response given by this student on a particular item. These factors include achieving motive (motivation), test anxiety, academic self-concept and attributive style.

Throughout the very large and ever-increasing volume of literature on the topic, test anxiety and motivation are deemed to be major factors contributing to test-score variance. Various models have been used to explain the link between test anxiety and academic achievement. Sarason (1984, p. 936), who views anxiety as 'self-preoccupation over the inability to respond adequately to the call', conceptualises test anxiety on four dimensions: worry, tension, test-irrelevant thinking and bodily symptoms.

According to Marsh (1990), highly motivated students are likely to agree strongly with the following statements: 'I see doing well in school as a sort of game, and I play to win.' 'I will work for top marks in a subject whether or not I like the subject.'

'I have a strong desire to do best in all of my studies.' 'I try to obtain high marks in all my subjects because of the advantage this gives me in competing with others when I leave school.'

Marsh's (1990) 20-item questionnaire for measuring 'school-subjects self-concept' includes items such as: 'People come to me for help in most school subjects.' 'If I work really hard I could be one of the best students in my year at school.' 'I learn things quickly in most school subjects.' 'I do well in tests in most school subjects.'

Some students attribute their success and failure to internal stable factors: 'I bombed out on that test because I have no talent.' Other students attribute their success and failure to external, unstable factors: 'I bombed out on that test because the teacher set stupid questions.' These two sets of students have different attributive styles; the former students have an internal locus of control, the latter an external locus of control.

When teachers and schools use assessment information, whether from international and national tests of generic skills, or from systemic tests of discipline-specific knowledge, they should not limit their explanations of low (or high) scores to teacher or school effect.

Features of the Assessment Process

The student–item interaction is also affected by features of the assessment process, which includes all those things that are experienced by the student as a result of decisions made by those who develop and administer the assessment instrument. The testing process (assessment under standardised conditions on an instrument that has been trialled beforehand) is obviously multifaceted, from what is put in front of the student to what the student is required to do, to the conditions under which the student is to function.

Features of the assessment process impose difficulty on the item that is not simply a function of its intrinsic difficulty (that is, nature of the cognitive task)—some concepts are, quite simply, 'hard' for most people. It could be a function of the way the test is designed, for example, format (multiple choice or extended writing) and mode (written or oral). Design-imposed difficulty exists and it affects different students in different ways. For example, Willingham and Cole (1997) note gender differences related to test format (multiple choice and free response); Stage (1994) notes gender differences in spatial ability (and its consequences for test design).

What Makes an Item Difficult?

Intrinsic difficulty and design-imposed difficulty are alluded to above. Together with the notion of self-imposed difficulty, these potential sources of empirical (statistical) difficulty provide a typology for explaining item difficulty (Matters, 1997).

A common question asked by teachers when examining aggregated data from standardised tests is: 'What made this multiple-choice item so difficult that only a small proportion of students chose the correct answer?' Setting aside the possibility

that the keyed response for marking was actually wrong, some of the questions in the set below might be useful in formulating an answer. The questions are composed for application to formats beyond multiple choice, to constructed response (active or passive) and extended response (as in a writing task or providing the solution to a substantial physics problem).

- What kind of thinking was involved: concrete, conceptual or personal?
- What abilities were required: verbal, numerical or spatial?
- What emphases were placed on the treatment of the stimulus material: Did the student need to absorb it, operate on it or transform it into something new?
- Is it possible that a student's (or the student group's) perception of success on the item was influenced by features of the stimulus material such as *context*?

The possibility that the context in which a test item is set imposes differential difficulty on students is of serious concern to some researchers investigating effects of the design of OECD's PISA (which by nature is context-bound, the context for the items being 'real life').

Design-Imposed Difficulty and PISA Results

A curious by-product of the release of comparative data from PISA (Thomson, Cresswell, & De Bortoli, 2004; Thomson & De Bortoli, 2008) is the almost-palpable performance anxiety at the level of participating countries and states. Even more curious is the not-infrequent spectacle, at conferences and other national and international gatherings, of countries defining themselves in terms of their PISA results. This phenomenon is observed from low- as well as high-performing countries.

The purpose of this short section is not to till the fertile ground of social, methodological and theoretical issues regarding PISA. The purpose is merely to tell a story that illustrates the explanatory power of the concept of design-imposed difficulty and, to a certain extent, self-imposed difficulty. For this I draw on Rochex's (2006) secondary and complementary analyses of the PISA 2000 literacy tests:

> Many of the PISA literacy tests required students to mobilise various fields of reference and various registers of resources and to combine and organise the elements that they could draw from these fields and registers into a hierarchy. The issue of hierarchy was all the more the case given that the goal of the PISA designers was to assess 'the skills to carry out tasks that belong to real-life situations', rather than specific knowledge, and that their themes were often close to the social and cultural references and experiences of the young people taking the test. (p. 185)

One of the conclusions of the study of students' methods (part of the larger study) was that, 'for a great number [of students], these methods varied more in relation to the texts and *contexts*, topics, and *type of tasks or question formats* than to their sole text treatment and reading and writing competencies—what was supposedly being assessed' (Rochex, 2006, p. 204) (my emphasis).

Rochex's finding has implications for the preparation of students for international surveys and also for national and state tests of generic or cross-curriculum skills,

where skills that have been developed through the experienced curriculum (the study of several academic subjects) are then tested in unfamiliar contexts.

Effective Use of Assessment Information

Policy makers and practitioners demand to know what works, to know when it works and for whom, and to know how it works and why. At the simplest level, getting to know *what works* comes from inference (and requires a plausible model for causality and a study of the data and their associations); getting to know w*hen it works* and *for whom* comes from generalisation; and getting to know *how it works* and *why* requires other methods.

Teachers and schools mostly want to know what students have achieved. Teachers, schools and policy analysts often want to know the conditions under which students or certain groups of students achieve. Responses to these demands can be found in the data through inference and generalisation. Responses to demands for explanations about the 'how' and 'why' can be sought in the fields of neurobiology, sociology and psychology.

Bialecki (2008, p. 91) describes how, based on data obtained through PISA surveys on literacy in 2000, 2003 and 2006, the distribution of low literacy has changed in Poland. Five factors were identified as contributing to differential performance on PISA literacy in Poland; two of those five factors were identified as changing after a targeted intervention (see Table 11.1).

Some of the variables are stable within an individual and some can be changed. The finding of interest is that, in response to intervention, it was possible to change student motivation and the literate environment (part of the milieu created by school life and home life). The significance of the literate environment seems to prove what many of us have always suspected about students whose milieu values the various ways in which the life of the mind manifests itself in everyday surroundings (books, images and so on). This finding has implications for countries or jurisdictions that are considering the possibility of joining the 'PISA club'.

Another example of how testing information can be used effectively is the examination, by teachers, of test data that illuminate students' misconceptions (some of which are classic). Because the possibility of having electrodes attached

Table 11.1 Factors influencing PISA literacy scores in Poland

Factor influencing PISA literacy scores, Poland	Classification according to framework in Fig. 11.1	Direction of change
Student ability	Background	↔
Student motivation	Psychological	↑
Parent social status	Background	↔
School attended	Background	↔
Literate environment	Milieu	↑

Source: Bialecki, 2008.

to the student's brain is not yet an option for obtaining more direct information about the student–item interaction, studying individual test responses seems to be a promising compromise as a tool for understanding errors in student reasoning. Other than students' misconceptions (or mere lack of knowledge), there is another oft-overlooked source of incorrect responses on a test—the instrument itself. Sometimes, no evidence of learning is to be found in student responses—not because there was no learning but because the items for bringing forth evidence of learning were flawed.

In the early 21st century climate of comparative test data, teachers, schools and even countries appear to be spending a disproportionate amount of time devising hypotheses to account for underperformance on assessment instruments such as PISA and national testing programs in literacy and numeracy, rather than having first studied the content and construct of the tests. If teachers and schools were to transfer some of that energy to a critique of the test items per se, they would be in a strong position to comment on the quality of the instrument, their own assessment/test development skills would be enhanced and they would be in a better position to prepare students for the test (a good test is worth teaching to).

None of the above is intended to undermine the importance of professional conversations about differential performance by country. International comparisons are seductive. Finland, a top scorer on PISA, is the target of interminable questioning about its success. Australians would be better off asking why it is that results from the Australian Capital Territory, one of the eight states and territories that make up the nation, are similar to those of Finland, and then leave it to policy makers in all countries to ponder the effects of highly trained subject-matter experts in the primary school and of promotion from one year level to the next that is not automatic.

Turn now to the proclivity to fixate on test data without ever querying the quality of the assessment instruments from which the data were generated. Conference papers and media reports are filled with references to information derived from test scores. Three examples of the thousands that exist are how different countries score on PISA, how different Australian states and territories score on national tests of literacy and numeracy and how bad the level of mathematics or science knowledge is in a certain place at a certain time.

It is quite extraordinary that there are so few, if any, conference papers and media reports that point out flawed items on high-stakes tests or query the key for a multiple-choice item or demand to know anything of post-test analyses. It is acknowledged that test development is a sophisticated industry circa 2008 and that test-development agencies have sophisticated quality assurance procedures. It may be the case that an infinitesimal proportion of flawed items appear on high-stakes tests around the world. It may be the case that the wrong option is never marked as correct on a high-stakes test anywhere in the world. It may be the case that post-test analyses always deliver acceptable values for vital parameters. What is surprising is that people, particularly students and teachers in a testing situation, usually challenge information that is not flattering to them, or attribute their lack of success to external factors such as the test itself. Cronbach (1988, p. 7), citing Campbell, declares that highlighting uncertainties can contribute to validity arguments and

that ‘a community should be disputatious’. Teachers and schools would do well by becoming part of the dialogue through evaluation and challenging of conclusions.

In order to illuminate a different approach to engaging with data generated by standardised tests, I draw attention to the fact that, while we religiously invoke the two main purposes of using assessment data (for learning and for reporting), we often forget the worthwhile learning experience that students get when they receive feedback from tests (not just learning about themselves, meta-cognition and so on, but learning to understand material that was originally not understood by them). For a teacher, the central purpose of using assessment information is to improve the learning of one or more particular students; that is, the individual teacher and the school take the students who come to them and seek to improve the learning of those students. Another purpose—pure intellectual curiosity about how students think—is not so prominent in today’s discourse.

I have sat through a 3-hour discussion about the functioning of a multiple-choice mathematics item on a nation-wide test. This item had been trialled before selection for the test. The lively discussion was not in English, although at regular intervals I was informed of the various hypotheses being devised to explain the high value for empirical difficulty: some plausible explanations included the use of vague language, verbal loading noted in mathematics testing, the non-parallel use of terminology in curriculum documentation and test item, the ambiguity in terms used from geometry (side versus edge; size versus volume), and even the possibility that the distracters were not tapping into classic misconceptions of students of this age in this domain. There had obviously not been a study of the variation in location of the item on the item–person map between trial and live administration. Nor was this information requested at any stage in the post-test discussion session. Mathematics and music are, arguably, two of the subjects in which one is most likely to be able to engage if the language being spoken is foreign, while the test item under discussion is highly visual or numerical. With great trepidation I ventured that I could not see how they had reached the ‘correct’ answer . . . and it transpired that there had been a clerical error in recording the keyed response. Was this discussion time wasted? No, for two reasons. First, there was the hard lesson learnt about transcription errors, which I will not labour here. Second, there was the sustained conversation, albeit for the wrong reasons, about how students think.

Items are relatively simple things compared with people, even though the mathematics of item analysis (Crocker & Algina, 1986; Hambleton, Swaminathan, & Rogers, 1991; Holland & Wainer, 1993) might create a different impression. But computer packages can give instant access to the world of item statistics and item-response modelling, and rules of thumb are composed within testing agencies on the use of information thus generated about items and students. It does little good to use a rule of thumb if a deeper understanding of its meaning could have led, instead, to the occasional (*and* profitable) breaking of the rule (for example, in selecting items on the basis of their trial statistics for inclusion in a test) . . . or, in the case of teachers and schools being provided with information about students’ test performance, to their being given an insightful reading of the data rather than being fobbed off by a confidently stated rule of thumb that had been applied.

What Information Is Worth Looking at?

It would take a text book to cover all the issues surrounding the use of assessment information. This section describes just two issues that teachers and schools need to look out for when presented with assessment information for their use.

This School Is More Successful than That School

Using information about school performance in ways that might damage individuals and organisations is an ethical issue.

Schools do not automatically increase the achievement level of their cohort of students over a given period of time to the same extent. That is, students at one school gain an additional advantage over students at another school. This relative advantage is known as 'value-added'.

The call for fair measures of school performance has generated statistical models with an emphasis on value-added (Goldstein, 2001; Kingsbury & Houser, 1997; Rowe, 2005). It is what the school has been able to add to the achievement of its cohort of students, given the ability of the students. Statisticians call it 'the residual' because it is that which is left over after they have taken student ability into account in their multiple-regression analyses. Thus, the residual is not just a measure of the influence of the school. Although there are measurement errors in its calculation, it is a more respected indicator of the net effect that schools have on student progress than a set of 'league tables'.

Through these value-added models, school or teacher effects are derived from complex analyses of limited datasets. The use of measures of 'value-addedness' is accompanied by serious difficulties in principle and in practice, not to mention the fact that the use of multi-level modelling creates a structural misalignment between the humane missions espoused by schools and the technocratic ways in which society increasingly measures the success of schools. League tables, for example, do not recognise a schools' success in adding value in *all* the main ways identified as critical to students' social and economic futures. They do not even recognise a schools' success in adding value in an academic way because they only indicate academic achievement at a point in time (when students 'graduate' from high school), thus assuming that all students were equivalent when they entered school. On the other hand, value-added measures, although restricted to academic achievement, do take account of ability.

Distasteful as these measures might be to some teachers and schools, 'we no longer have the luxury as a society to view comparisons [between schools] as invidious' (Allen, 2007, p. 12). If we accept the political reality of comparisons of school performance, we then encounter another problem—a dearth of sophisticated methods for making the comparisons. If we use the available technology to manipulate large datasets (for example, for cluster analysis), we then impose clusters on the data. In some Australian states, each cluster comprises the so-called 'like schools'. Allen (2007, p. 11), with a dash of acerbity, writes what many have only thought:

'we can no longer restrict comparisons to "like schools" because it is abundantly clear that it is not the schools that are alike [in fact the schools are simply not alike]'.

Female Performance Is Better than Male Performance

Using only means and standard deviations (or just means) for reporting differences in achievement between subgroups of the population has limitations that need to be recognised.

Reports in the media about gender differences tend to focus on mean performance when comparing the test results of females and males when, in fact, there may be much more to say than that one group or other has a higher average. It is possible, for example, for one group to have a higher mean but for the other group to have more of the higher results. This sort of relationship can be captured in a single picture, the Q–Q plot as in Fig. 11.2. It is a plot of the quantiles of the male achievement distribution against the corresponding quantiles of the female achievement distribution. If the points lie along the straight line, $y = x$, the distribution of male achievement matches the distribution of female achievement. A segment of the points above the straight line means that the boys in that part of their achievement distribution did better than the girls in the corresponding part of their achievement distribution.

Figure 11.2 compares the overall achievement indicators, from which tertiary entrance ranks were determined for males and females in Queensland, Australia, in 1988. The graph shows that the males in the top third do better than the females in the top third, whereas the males right at the bottom do much worse than the females. The segment of points below the straight line ($y = x$) covers most of the range of achievement plotted. This tells us that the girls are ahead of the boys over most of

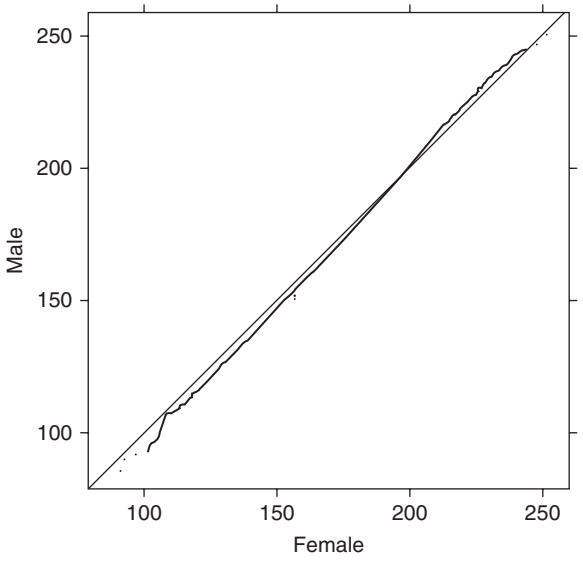

Fig. 11.2 Q–Q plot of overall achievement, by gender, Queensland, 1998 (Source: Matters, Allen, Gray, & Pitman, 1999, p. 296)

the range *but* the boys are ahead at the top. In common parlance, the boys both shine most at the top and 'bomb' out best at the bottom.

Had only the summary statistics been presented, the lingering piece of information would have been that the female group had the higher average score—which could lead to newspaper article headlines of the style 'Girls outperform the boys', when in fact the boys have more of the higher results.

What Has Changed?

Towards the end of the 20th century, Linn (1989, p. 1) listed the advent of hugely sophisticated methods for *measuring* student performance as one of the three most significant changes in educational measurement over the preceding 18 years. Matters (2006, p. 5) argues that over the last decade of the 20th century (the dawning of a new accountability age), the locus of interest moved to the practice of *using* information from the assessment process. I now argue that, at the beginning of the 21st century, when results from international tests and surveys stimulate educational discussion and debate, the significance of student responses at the item level is in the spotlight. If this is the case then teachers and schools should demand streamlined publications containing information about items and students on external standardised tests (and not just PISA).

There is no point in using assessment information for any purpose unless the assessment instrument is good. Mapping backwards and forwards from Bennett's (2006) take on Mislevy, Almond, and Lukas (2003) produces a dependency sequence in two directions: one, useful assessment information comes from good design (design that proceeds after attending to the precursors); and, two, good assessment tasks come from paying attention to what is going to happen with the data on student achievement.

There is no point in using assessment information if the user does not understand the form and purpose of assessment and the act of assessing (whichever paradigm dominates—judging or measuring), including the nature of admissible evidence of student learning. There is no point in using assessment information if the user is not aware of the psychometric underpinnings of assessment or does not possess the skills necessary for interpretation of student achievement data. There is no point in accepting the total score on a collection of test items as a measure of the underlying construct if there is any doubt at all about the properties of individual items in the test.

Conclusion

This chapter is part of a collection of writings around the theme 'assessment issues for the 21st century'. It documents differences in the use of assessment information between the late 20th century and the 21st century thus far, and it underlines schools' and teachers' use of assessment information, especially information from external

standardised tests. In the case of well-designed standardised tests, the product (student responses) gives information not only about what was learnt and how well it was learnt but also about what was not learnt and hints as to why this might be so. This chapter also attempts to convince teachers and schools to study tests and student responses at the item level in order to confirm the existence of evidence to support the proposed construct interpretation of test scores.

When Socrates, on trial for heresy, said 'the life which is unexamined is not worth living', he was not referring to the need for public examinations or standardised tests; but if he did say just that in today's educational environment, it is unlikely he would be put to death for it. The role of assessment, and of assessment information, in educational debate and policy in the early 21st century is an extremely powerful one. This chapter contends that this role can be justified only if two conditions (at least) are met: that the assessment itself is of sufficient strength and quality to support the uses to which it will be put, and that the users of the assessment data—the analysts, the teachers, the administrators, the policy makers—have sufficient expertise and imagination to see beyond the rules of thumb and piece together the true underlying story (whether the story in question is about a student underachieving in one subject, or a country outperforming Finland on international standardised tests).

Then it is possible to make the links (forward and backward) between the three points of presage, process and product, in ways that maximise the usefulness of the information obtained not only about the tangible product but also the process (the intangible student–item interaction). Without this level of rigour and expertise being applied to the assessment on which so much today is based, we could adapt what Socrates said, and say that the assessment which is unexamined is not worth using.

Theoretical and Methodological Framings

Paradigms: Measurement Versus Judgment

Educational assessment is the collection of information about student learning in numerous ways for two main purposes: for feeding back into the learning process and/or for reporting to various audiences. Evidence of student learning is obtained in response to assessment instruments. Decisions about the extent of that learning are coded as assessment results. Two paradigms operate: *measuring* how much of a certain quality (single underlying dimension) is evidenced in student responses; and *judging* what the evidence says about what the student has learnt and how well.

The psychometric model that 'observed score = true score + error' suits notions of reliability and validity for multiple-choice testing. Assumptions of the true-score model do not readily suit notions of reliability and validity for testing in open-ended response modes and do not at all suit notions of validity and reliability for school-based assessment. Here, assumptions of the true-score model do not hold, in particular, assumptions about infinite populations, about markers, items and tasks being sampled at random from a universe of markers, items and tasks, and about identical and independent Gaussian distributions. In many school settings, split-half

reliability estimates are not possible and the practice of inter-rater agreement studies is beyond the resources of most schools. According to Moss (1992, 1994) the epistemological and ethical purposes served by reliability can be broadened to include the practice of contextualised judgment. One of her three warrants for reliability is the privileging of contextualised (teacher) judgments. This involves the use of a criteria and standards schema against which teachers judge the quality of student work.

Glossary

Constructed response Refers to assessment items in which students are required to produce a short answer (as opposed to, for example, writing an essay, doing a project or selecting the correct response from a list of options). Responses might involve writing a paragraph of exposition or explanation, performing a calculation, constructing a graph, compiling a table, or producing a sketch or drawing.

> *Active constructed-response items* require the candidate/student to take the stimulus material and do something with it, as in calculating and summarising, or even transforming it into something new, as in composing a poem or devising a plan.
> *Passive constructed-response items* require the candidate/student to treat the stimulus material in a reflexive way, to absorb it as in interpretation or in searching/locating in order to quote extracts

Empirical (statistical) difficulty In a multiple-choice test, the facility index (*f*) for an item is defined as the proportion (percentage) of candidates giving the correct response. The lower the value of *f,* the more difficult the item experientially

Item-response modelling/item-response theory Item-response theory combines psychology and mathematics in determining the probability p that an examinee with ability θ correctly answers an item. Modern test theory (after Rasch) estimates item parameters and person ability, placing items and students on the same scale

Item statistics In terms of classical test theory, key statistics for an item (from trial test or 'live' test) are:

> *Difficulty* (see *empirical difficulty* and *facility index*).
> *Discrimination* (usually the point biserial correlation, which is a form of product–moment correlation, between item score and test score)

Quantile The quantile of a distribution of values is a number that indicates the proportion of the values that are less than or equal to that value. For example, the 0.75 quantile of a variable is a value below which 75 per cent of the values of the variable fall

Socio-economic status A measure of an individual's or group's position in the social order in terms of income, occupation, educational attainment, wealth, etc.

Standardised testing Involves all, or a wide cross-section of, students across a jurisdiction, of the same year or age sitting (versions of) the same test under the same

conditions and, usually at the same time, with results being reported in a common format (on the same scale and/or according to a commonly applied marking scheme)

Test/test item A published instrument constructed by persons technically trained in mental testing and statistical methods. Its *items* have been thoroughly tried out beforehand, and the test is accompanied by norms or standards of performance that enable the tester to interpret how far a student's score or mark is superior or inferior to those of other similar students

Underlying attribute The theoretical, intangible quality or trait that allows for individual differences in that quality or trait to be measured

References

American Education Research Association, & American Psychological Association & National Council on Measurement of Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Allen, J. R. (2007). *Fair measures of school performance*. Paper presented at the 2007 ACACA Annual Conference, Melbourne.

Bennett, R. E. (2006). Foreword. In G. N. Matters, *Using data to support learning in schools: Students, teachers, systems*. Camberwell: ACER Press.

Bialecki, I. (2008). *Assessment measurement and evaluation of literacy levels and of basic competencies in Poland*. Paper delivered at UNESCO Regional Conference (Europe) in Support of Global Literacy, Baku, Azerbaijan.

Biggs, J. B. (1993). From theory to practice: A cognitive systems approach. *Higher Education Research and Development*, *12*, 73–86.

Biggs, J. B. (1999). *Teaching for quality learning at university*. Buckingham: SRHE & Open University Press.

Biggs, J. B., & Moore, P. J. (1993). *The process of learning*. (3rd ed.). New York: Prentice Hall.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity.* Hillsdale, NJ: Erlbaum.

Ericson, D. P., & Ellett, F. S. (2002). The question of the student in educational reform. *Educational Policy Analysis Archives*, *10*(31), Retrieved June 19, 2008, from <http://epaa.asu.edu/epaa/v10n31/>.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: Scope and limitations. *British Educational Research Journal*, *27*(4), 433–442.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hattie, J. A. C. (2005). *What is the nature of evidence that makes a difference to learning?* Paper delivered at the ACER Conference, Melbourne.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kingsbury, G. G., & Houser, R. (1997). Using data from a level testing system to change a school district. In J. O'Reilly (Ed.), *The Rasch tiger ten years later: Using IRT techniques to measure achievement in schools.* Chicago: National Association of Test Directors.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd edn.). Washington, DC: The American Council on Education and the National Council on Measurement in Education.

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review*, *2*(2), 77–172.

Matters, G. N. (1997). *Are Australian boys underachieving?* Paper presented at 23rd annual conference of the International Association for Assessment in Education. Durban, South Africa.
Matters, G. N. (2006). Using data to support learning in schools: Students, teachers, systems. Camberwell: ACER Press.
Matters, G. N., Allen, J. R., Gray, K. R., & Pitman, J. A. (1999). Can we tell the difference and does it matter? Differences in achievement between girls and boys in Australian senior secondary education. *The Curriculum Journal*, *10*(2), 283–302.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centered Design*. Retrieved June 19, 2008, from <http://www.ets.org/Media/Research/pdf/RR-03-16.pdf>.
Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229–258.
Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5–12.
OECD (Organisation for Economic Co-operation and Development). (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: Author.
Rochex, J. Y. (2006). Social, methodological, and theoretical issues regarding assessment: lessons from a secondary analysis of PISA 2000 Literacy Tests. *Review of Research in Education*, *30*(1), 163–212.
Rowe, K. J. (2006). *Evidence for the kinds of feedback data that support both student and teacher learning.* Paper delivered at the ACER Conference, Melbourne.
Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*(4), 929–938.
Stage, C. (1994). *Gender differences on the SweSAT: A review of studies since 1975*. Department of Educational Measurement, Umeå University, EM No. 7.
Thomson, S., & De Bortoli, L. (2008). *Exploring scientific literacy: How Australia measures up. The PISA 2006 survey of students' scientific, reading and mathematical literacy skills.* Camberwell: ACER Press.
Thomson, S., Cresswell, J., & De Bortoli, L. (2004). *Facing the future: A focus on mathematical literacy among Australian 15-year-old students in PISA 2003.* Camberwell: ACER Press.
Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Princeton, NJ: Lawrence Erlbaum.