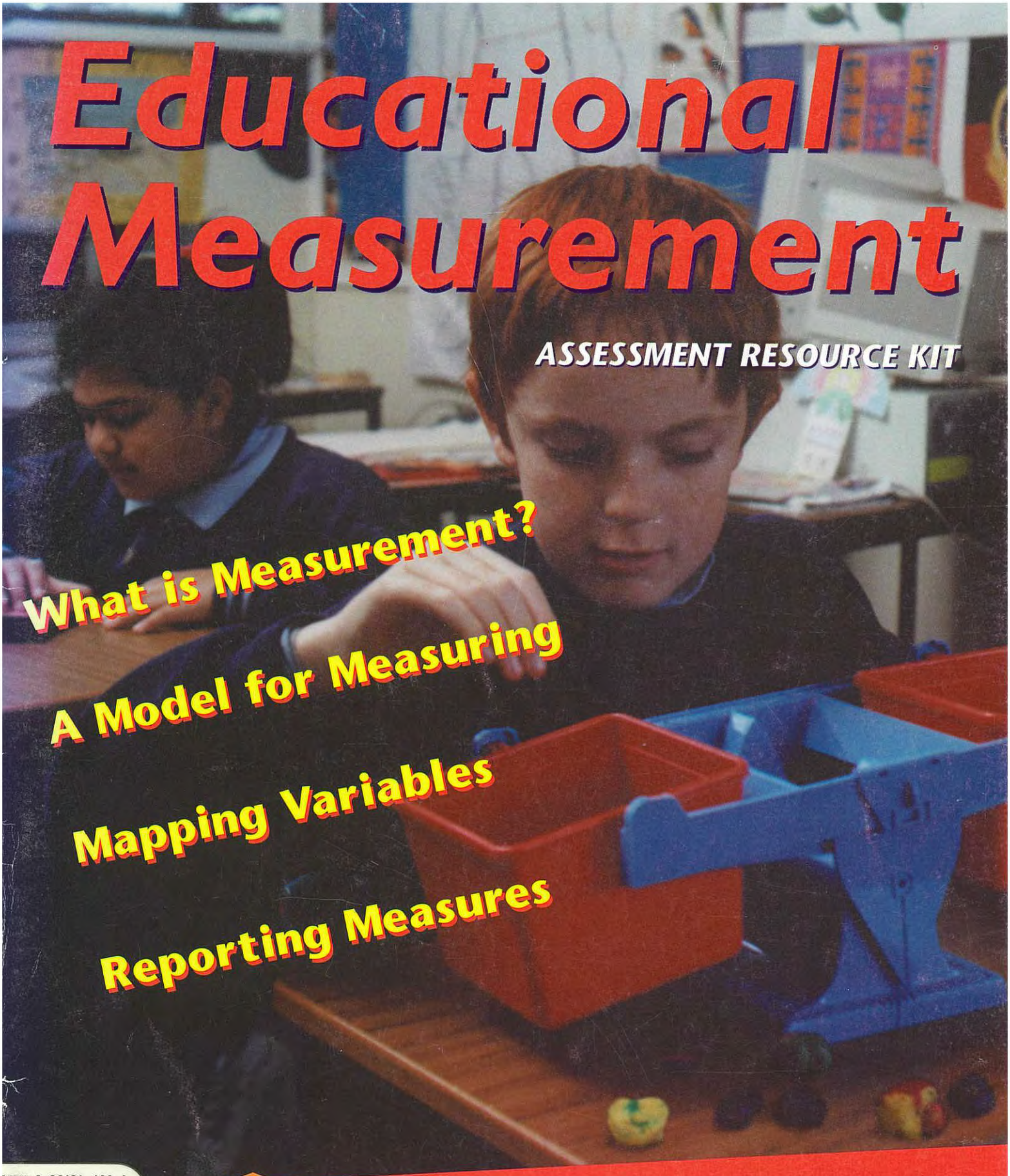
# Educational Measurement

ASSESSMENT RESOURCE KIT



What is Measurement?

A Model for Measuring

Mapping Variables

Reporting Measures

ISBN 0-86431-400-0

780864 314000

A108ARK

ARK

Geoff N Masters

# contents




ACER

Australian Council for Educational Research
19 Prospect Hill Road (Private Bag 55)
Camberwell Victoria AUSTRALIA 3124

# Glossary

*ability*
: a generic term for an individual's location on a measurement variable (ie, a parameter to be estimated)

*calibration*
: the process of estimating the difficulties of test items from students' responses to them

*computer adaptive testing*
: a process by which items are selected for administration one at a time from an item bank on the basis of a student's performance on preceding items

*criterion referencing*
: the process of interpreting individuals' test performances in terms of specified learning objectives (sometimes used to describe the interpretation of test performances in terms of a specified performance standard)

*described proficiency scale*
: a measurement variable/scale described in terms of the knowledge, skills, understandings, attitudes or values typically observed at various locations along that scale

*dichotomous scoring*
: the scoring of individuals' item responses in two categories only (usually right/wrong)

*differential item functioning*
: the observation that an item is atypically easier or harder for one group of students than for another (eg, unusually difficult for females)

*difficulty*
: an item's location on a measurement variable (a parameter to be estimated)

*equating*
: a statistical process that converts scores on different tests to the same scale, allowing them to be compared directly

*fit analysis*
: the statistical analysis of how well responses to an item (or by a person) match the expectations of a measurement model

*item*
: a test question or task

*item bank*
: a collection of test items calibrated on the same measurement variable

*item bias*
: the observation that an item is atypically easier or harder for one group of students than for another (eg, unusually difficult for females)

*link items*
: items shared by two or more tests, allowing those tests to be equated

*logit*
: a unit of measurement

*measurement*
: the process of estimating students' locations (abilities) on a measurement variable from their responses to a set of items

*measurement error*
: an indication of the uncertainty associated with the estimate of a student's ability

*norm referencing*
: the process of interpreting individuals' test performances in terms of the performances of a relevant reference group (eg, students of the same age)

*objectivity*
: a characteristic of a measurement system enabling measures to be compared without regard to the particulars of the tasks used or judges involved

*outcomes*
: the results of learning (eg, knowledge, skills, understandings, attitudes, values)

*partial credit*
: the scoring of individuals' item responses in several ordered categories

*performance standards*
: levels of ability set as targets or requirements for particular purposes (usually operationalised as 'cut-scores')

*progress map*
: (see described proficiency scale)

*Rasch model*
: a measurement model capable of providing objective measures in a defined unit of measurement

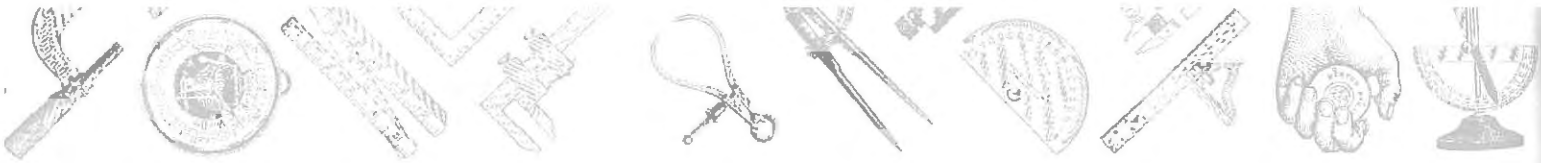
*ratings*
: judgements of individuals' performances or responses made in terms of a set of ordered categories (ie, a rating scale)

*standard setting*
: the process of setting a performance standard

*unidimensionality*
: an idealised state in which individuals' responses to a set of items are governed only by the variable those items are designed to measure

*unit of measurement*
: a constant amount of a continuous variable that can be repeated and counted

*variable*
: something that varies – in this context, an ability (attitude etc) that can be conceptualised as varying along a continuum



# what is 'measurement' ?

## Conceptualising Variables

In life, the most powerful ideas are the simplest. Many areas of human endeavour, including science and religion, involve a search for simple unifying ideas that offer the most parsimonious explanations for the widest variety of human experience.

Early in human history, we found ourselves surrounded by objects of impossible complexity. To make sense of the world we found it useful, and probably necessary, to ignore this complexity and to invent simple ways of thinking about and describing the objects around us. One useful strategy was to focus on particular ways in which objects differed.

The concepts of 'big' and 'small' provided an especially useful distinction. Bigness was an idea that allowed us to ignore the myriad other ways in which objects differed – including colour, shape and texture – and to focus on just *one* feature of an object: its bigness. The abstract notion of 'bigness' was a powerful idea because it could be used in describing objects as different as rivers, animals, rocks and trees.

For much of our history, the concept of 'bigness' no doubt served us well. But as we made more detailed observations of objects, and as we reflected on those observations, we found it useful to distinguish size from weight, even though size and weight usually were closely related. And, as we grappled with our experience that larger objects were not always heavier, we introduced the more sophisticated concepts of density and specific gravity.

Each of these ideas provided a way of focusing on just *one* way in which objects differed at a time, and so provided a tool for dealing with the otherwise unmanageable complexity of the world around us. Bigness, weight, length, volume and density were just some of our ideas for describing the ways in which objects varied; other 'variables' included hardness, temperature, inertia, speed, acceleration, malleability, and momentum. As our understandings improved and our observations became more sophisticated, we found it useful to invent new variables subtly distinguished from existing variables: for example, to distinguish mass from weight, velocity from speed, and temperature from heat.

The advantage of a variable was that it allowed us to set aside – at least temporarily – the very complex ways in which objects differed, and to see objects through just one lens at a time. For example, objects could be placed in a single order of increasing weight, regardless of their varying shapes, colours, surface areas, volumes, and temperatures. The weight 'lens' allowed us to see objects on just one of an essentially infinite number of possible dimensions.

We sometimes wondered whether we had invented these variables or simply discovered them. Was the concept of momentum a human invention, or was momentum 'discovered'? Certainly, it was a human decision to focus attention on specific aspects of variability in the world around us and to work to clarify and operationalise variables. The painstaking and relatively

recent work of Anders Celsius (1701–44) and Gabriel Fahrenheit (1686–1736) to develop a useful working definition of temperature was testament to that. On the other hand, the variables we developed were intended to represent 'real' differences among objects. Ultimately, the question of whether variables were discovered or invented was of limited philosophical interest: the important question about a variable was whether it was useful in practice.

dimension
[di(s) — apart;
metiri — to measure]:

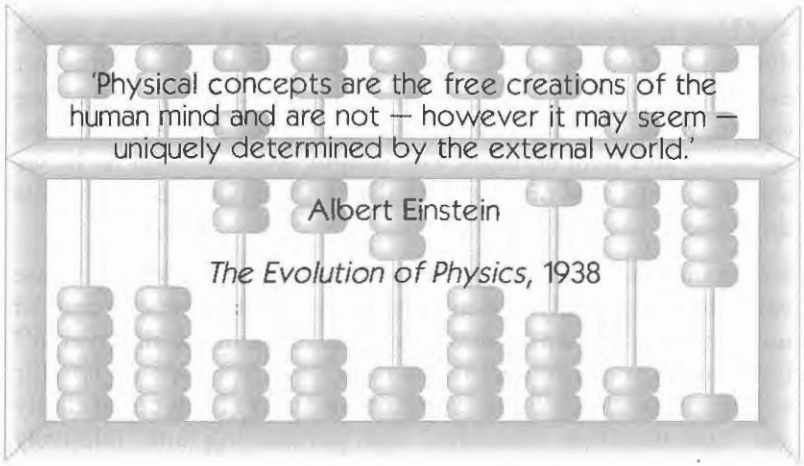
separated out for measurement

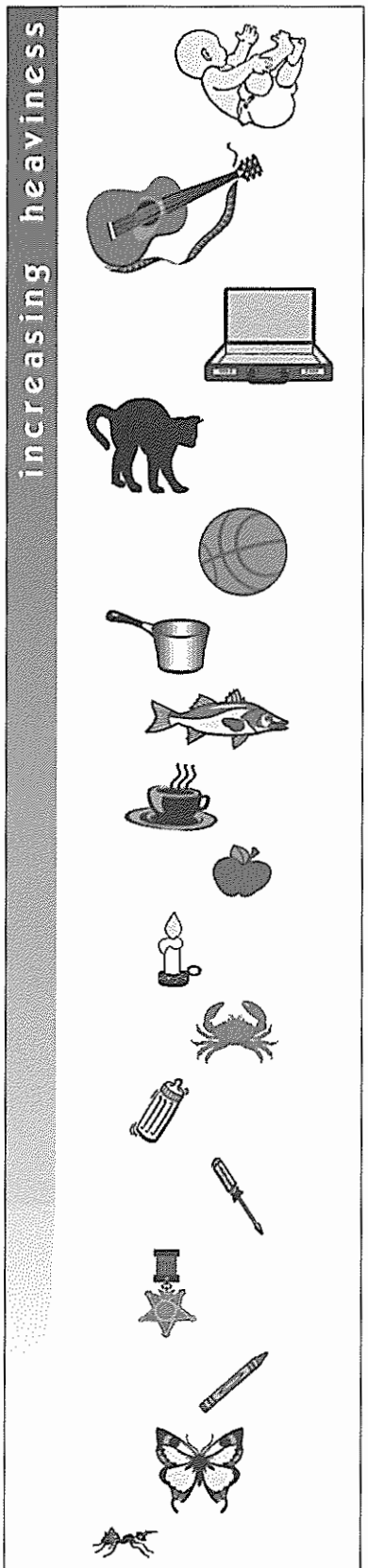
## Human Variability

But it was not only inanimate objects that were impossibly complex; people were too. Again, a strategy for dealing with this complexity was to focus on particular ways in which people varied. Some humans were faster runners than others, some had greater strength, some were better hunters, more graceful dancers, superior warriors, more skilled craftsmen, wiser teachers, more compassionate counsellors, more comical entertainers, greater orators. The list of dimensions on which humans could be compared was unending, and the language we developed to describe this variability was vast and impressive.

In dealing with human complexity, our decision to focus on one aspect of variability at a time was at least as important as it was in dealing with the complexity of inanimate objects. To select the best person to lead the hunting party it was desirable to focus on individuals' prowess as hunters and to recognise that the best hunter was not necessarily the most entertaining dancer around the campfire or the best storyteller in the group. There were times when our very existence depended on clarity about the relative strengths and weaknesses of fellow human beings.

The decision to pay attention to *one* aspect of variability at a time also was important when it came to monitoring the development of skills, understandings, attitudes and values in the young. As adults, we sought to develop different kinds of abilities in children, including skills in hunting, dancing, reading, writing, storytelling, making and using weapons and tools, constructing dwellings, and preparing food. We also sought to develop children's knowledge of local geography, flora and fauna, and their understandings of tribal customs and rituals, religious ceremonies, and oral history. To monitor children's progress towards mature, wise, well-rounded adults, we often found it






The weight variable can be visualised as a continuum of increasing heaviness.

convenient to focus on just one aspect of their development at a time.

We sometimes wondered whether the variables we used to deal with the complexity of human behaviour were 'real' in the sense that temperature and weight were 'real'. Did children *really* differ in reading ability? Were differences in children's reading abilities 'real' in the sense that differences in objects' potential energy or momentum were 'real'?

Once again, the important question was whether a variable such as reading ability was useful in practice. Common experience suggested that children *did* differ in their reading abilities and that individuals' reading abilities did develop over time. But was the idea of a variable of increasing reading competence supported by closer observations of reading behaviour? Did this idea help us to understand and promote reading development? As with all variables, the most important question about dimensions of human variability was whether they were helpful in dealing with the complexities of human experience.

In summary, our decision to focus attention on one aspect of variability at a time was a significant breakthrough in the management of complexity. The conceptualisation of variables was our first step towards measurement.

## Inventing Units

The second step towards measurement was the invention of units representing equal amounts of the variable being measured. Important human progress in counting units was made in relation to the most intangible of variables: time.

Time, unlike other variables such as length and weight, could not be manipulated and was much more difficult to conceptualise. But, amazingly, man found himself living inside a giant clock. By carefully inspecting the rhythmical ticking of the clock's mechanism, man learned how to measure time by counting units of time.

The regular rotation of the Earth on its axis marked out equal amounts of time and provided humans with a basic unit of measurement: the day. By counting days, we were able to replace qualitative descriptions of time ('a long time ago') with quantitative descriptions ('five days ago'). This was the second requirement for measurement: a unit of measurement. A unit was a fixed amount of a variable that could be repeated without modification and counted. The invention of units allowed the question *how much?* to be answered by counting *how many* units.

The regular revolution of the moon around the Earth provided a larger unit of time, the 'moon' or lunar month. And the regular revolution of the Earth around the sun led to the seasons and a still larger unit, the year. The motion of these heavenly bodies provided us with an instrument for marking off equal amounts of time and taught us that units could be combined to form larger units, or subdivided to form still smaller units (hours, minutes, seconds).

Ancient civilisations created ways of tabulating their measurements of time in calendars chiselled in stone, and used moving shadows to invent units smaller than the day. By observing the rhythmical motion of the giant clock in which we lived, humans probably developed a sophistication in the measurement of time before we developed a similar sophistication in the measurement of more tangible variables such as length, weight and temperature.

The invention of units of measurement was equally crucial to accurate communication about distances. In man's early history, 'a long way' became '2-days walk', again allowing the question *how much?* to be answered by counting *how many* units. For shorter distances, we counted paces. One thousand paces we called a mile (mil). Other units of length we defined in terms of parts of the body – the foot, cubit (length of forearm), hand – or in terms of objects that could be carried and placed end to end: the chain; the link (1/100 of a chain); the rod, perch or pole (a pole); and the yard (a stick).

Our recent and continuing use of many of these units is a reminder of how recently we mastered the measurement of length. The same is true of the units we use to measure some other variables (eg, 'stones' to measure weight). And still other units were invented so recently that we know the names of their inventors (eg, Celsius and Fahrenheit).

*biblical measures*

And God said to Noah, '...This is how you are to make it: the length of the ark 300 cubits, its breadth 50 cubits, and its height 30 cubits.'

*Genesis 6:15*

Ephron answered Abraham, 'My lord, listen to me; a piece of land worth 400 shekels of silver, what is that between you and me?' ... and Abraham weighed out for Ephron the silver he had named ..., 400 shekels of silver, according to the weights current among the merchants.

*Genesis 23:15*

## Pursuing Objectivity

The invention of units such as paces, feet, spans, cubits, chains, stones, rods and poles which could be repeated without modification provided humans with instruments for measuring. However, an important question in making measurements was whether different instruments provided numerically equivalent measures of the same object.

If two instruments did not provide numerically equivalent measures, then one possibility was that they were not calibrated in the same unit. It was one thing to agree on the use of a foot to measure length, but whose foot? What if my stone was heavier than yours? What if your chain was longer than mine? A fundamental requirement for useful measurement was that the resulting measures had to be independent of the measuring instrument and of the person doing the measuring: in other words, they had to be *objective.*

To achieve this kind of objectivity, it was necessary to establish and share common, or standard, units of measurement. For example, in 1790 it was agreed to measure length in terms of a 'metre', defined as one ten-millionth of the distance from the North Pole to the Equator. After the 1875 Treaty of the Metre, a metre was re-defined as the length of a platinum-iridium bar kept at the International Bureau of Weights and Measures near Paris, and from 1983 a metre was defined as the distance travelled by light in a vacuum in 1/299,792,458 of a second. All measuring sticks marked out in metres and centimetres were calibrated against this standard unit. Bureaus of weights and measures were established to ensure that standards were maintained and that instruments were calibrated accurately against standard units. In this way,

And Joseph stored up grain in great abundance, like the sand of the sea, until he ceased to measure it, for it could not be measured.

*Genesis 41:49*

## Maintaining standards

Whereas Edward Masters, Thomas Draper, Henry Chesheire and Margaret Ball stand severally presented in this court for keeping and using unlawful strikes*, it is therefore this day ordered by this court that John Stratford, esquire, and Thomas Corbyn, esquire, shall be and are hereby desired to examine whether the said several persons have caused their strikes to be made equal to the brasen standard provided by the lord of the manor of Atherston and to certify their doings to this court at the next General Sessions of the Peace.

*Warwickshire, Epiphany 1673*

**strike = instrument for measuring grain?*

measures could be compared directly from instrument to instrument – an essential requirement for accurate communication and for the successful conduct of commerce, science and industry.

If two instruments did not provide numerically equivalent measures, then a second, more serious possibility was that they were not providing measures of the same variable. The simplest indication of this problem was when two instruments produced significantly different orderings of a set of objects.

For example, two measuring sticks, one calibrated in centimetres, the other calibrated in inches, provided different numerical measures of an object. But when a number of objects were measured in both inches and centimetres and the measures in inches were plotted against the measures in centimetres, the resulting points approximated a straight line (and with no measurement error, would have formed a perfect straight line). In other words, the two measuring sticks provided *consistent* measures of length.

However, if on one instrument Object A was measured to be significantly greater than Object B, but on a second instrument Object B was measured to be significantly greater than Object A, then that would be evidence of a basic inconsistency. What should we conclude about the relative standings of Objects A and B on our variable?

A fundamental requirement for measurement was that it should not matter which instrument was used, or who was doing the measuring (ie, the requirement of objectivity/impartiality). Only if different instruments provided consistent measurements was it possible to achieve this kind of objectivity in our measures.

## In Summary

Measurement is one of mankind's most powerful and significant inventions.

Measurement begins with the decision to pay attention to only one way in which objects or persons differ. This decision to focus on just one aspect of variability allows objects to be conceptualised as having a single order on a variable ('dimension'). The conceptualisation of a variable as a continuum of increasing amounts is the first step towards measurement.

The second step towards measurement is the invention of a unit. A unit is an amount of a variable that can be repeated without modification and counted. The use of a unit ensures that equal numerical differences represent equal amounts of the variable.

The third and final step is to ensure objectivity in measurement. Measures are objective when they do not depend on a knowledge of the particular instrument used to obtain them, or of the person involved in the measuring process. The test of objectivity is whether equivalent numerical measures are obtained with different instruments and with different persons doing the measuring.

# aspiring to measure

## Educational Variables

In educational settings it is common to separate out and pay attention to one aspect of a student's development at a time.

When a teacher seeks to establish the stage a student has reached in his or her learning, to monitor that student's progress over time, or to make decisions about the most appropriate kinds of learning experiences for individuals, these questions usually are addressed in relation to *one* area of learning at a time. For example, it is usual to assess a child's attainment in numerical reasoning separately from the many other dimensions along which that child might be progressing (such as reading, writing, and spoken language), even though those aspects of development may be related.

Most educational variables can be conceptualised as aspects of learning in which students make progress over a number of years. Reading is an example. Reading begins in early childhood, but continues to develop through the primary years as children develop skills in extracting increasingly subtle meanings from increasingly complex texts. And, for most children, reading development does not stop there: it continues into the secondary years.

Teachers and educational administrators use measures of student progress and attainment for a wide variety of purposes.

Measures on educational variables are sought whenever there is a desire to ensure that limited places in educational programs are offered to those who are most deserving and best able to benefit from them. For example, places in medical schools are limited because of the costs of providing medical programs and because of the limited need for medical practitioners in the community. Medical schools seek to ensure that places are offered to applicants on the basis of their likely success in medical school and, where possible, on the extent to which applicants appear suited to subsequent medical practice. To allocate places fairly, medical schools go to some trouble to identify and measure relevant attributes of applicants. Universities and schools offering scholarships on the basis of academic merit similarly go to some trouble to identify and measure candidates on appropriate dimensions of achievement.

Measures of educational achievement and competence are sought at the completion of education and training programs. Has the student achieved a sufficient level of understanding and knowledge by the end of a course of instruction to satisfy the objectives of that course? Has the student achieved a sufficient level of competence to be allowed to practise (eg, as an accountant, a lawyer, a paediatrician, an airline pilot)? Decisions of this kind usually are made by first identifying the areas of knowledge, skill and understanding in which some minimum level of competence must be demonstrated, and by then measuring candidates' levels of competence or achievement in each of those areas.

Measures of educational achievement also are required to investigate ways of improving student learning: for example, to evaluate the impact of particular educational initiatives, to compare the effectiveness of different ways of structuring and managing educational delivery, and to identify the most effective teaching strategies and most cost-effective ways of lifting the achievements of under-achieving sections of the student population. Most educational research, including the evaluation of educational programs, depends on reliable measures of aspects of student learning. Some of the most informative research studies track student progress on one or more variables over a number of years (ie, longitudinal studies of progress).

The intention to separate out and measure variables in education is made explicit in the construction and use of educational tests. The intention to obtain only *one* test score for each student so that all students can be placed in a single score order reflects the intention to measure students on just one variable, and is called the intention of unidimensionality. On such a test, higher scores are intended to represent more of the variable that the test is designed to measure, and lower scores are intended to represent less. The use of an educational test to provide just one order of students along an educational variable is identical in principle to the intention to order objects along a single variable of increasing heaviness (see page 3).

The intention to obtain only *one* score for each student so that all students can be placed in a *single* score order is known as the intention of unidimensionality.

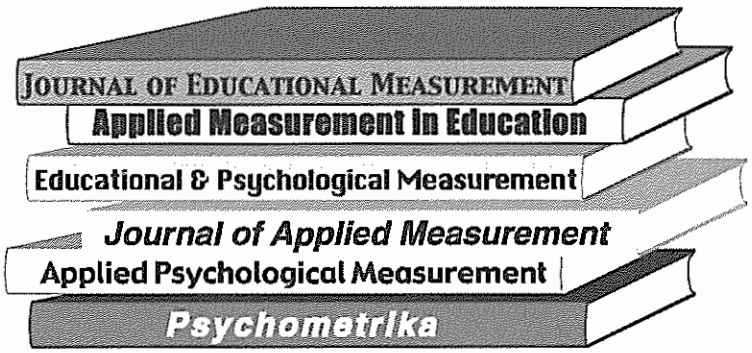
Occasionally, tests are constructed with the intention not of providing one score, but of providing several scores. For example, a test of reasoning might be constructed with the intention of obtaining both a verbal reasoning score and a quantitative reasoning score for each student. Or a mathematics achievement test might be constructed to provide separate scores in Number, Measurement and Space. Tests of this kind are really composite tests. The set of verbal reasoning items constitutes one measuring instrument; the set of quantitative reasoning items constitutes another. The fact that both sets of items are administered in the same test sitting is simply an administrative convenience.

Not every set of questions is constructed with the intention that the questions will form a measuring instrument. For example, some questionnaires are constructed with the intention of reporting responses to each question separately, but with no intention of combining responses across questions (eg, How many hours on average do you spend watching television each day? What type of book or magazine do you most like to read?). Questions of this kind may be asked not because they are intended to provide evidence about the same underlying variable, but because there is an interest in how some population of students responds to each question separately. The best check on whether a set of questions is intended to form a measuring instrument is to establish whether the writer intends to combine responses to obtain a single score for each student.

The development of every measuring instrument begins with the concept of a variable. The table on page 8 shows some of the many hundreds of variables listed in the *Mental Measurements Yearbook* for which measuring instruments (tests and questionnaires) have been constructed.

abstract reasoning
achievement orientation
adding fractions
adventurousness
aggression
altruism
anxiety
arithmetic ability
artistic expression
asocial behaviour
assertiveness
attentiveness
attitude toward mathematics
auditory perception
body satisfaction
business judgement
cautiousness
clerical accuracy
cognition of semantic relations
collaborating
communicating information
competitiveness
comprehending dialogues
computer programming
conscientiousness
copying shapes
creativity
critical thinking
decision making
depression
dexterity
division of decimals
ego strength
emotional resilience
empathy
expressiveness
extroversion
fine motor function
following directions
goal orientation
gross motor function
handwriting
hyperactivity
impulsiveness
interpersonal competency
intuitive thinking
kindergarten readiness
language comprehension
leadership potential
letter recognition
life satisfaction
listening comprehension
manual dexterity
mathematical understanding
memory for sentences
narrative writing
nutrition knowledge
oral reading
perception of objects in space
personal self-care
physical prowess
pitch discrimination
problem-solving ability
proofreading
reading ability
reasoning
relationship identification
sales comprehension
school adjustment
school readiness
self-confidence
sociability
spelling
stress tolerance
tactile differentiation
typing speed
verbal reasoning
visual discrimination
vocabulary knowledge
written expression

The intention underlying each of these instruments – and many others reported in the educational and psychological measurement literature – is to assemble a set of items capable of providing evidence about the variable of interest, and then to combine responses to those items to obtain measures of the variable. This intention raises the question of whether the set of items assembled to measure any given variable work together to form a useful measuring instrument.




## Equal Intervals?

When a student takes a test, the outcome is a test score, intended as a measure of the variable that the test was designed to measure. Test scores provide a single order of test takers – from the lowest scorer (the person who answers fewest items correctly or who agrees with fewest statements on a questionnaire) to the highest scorer. Because scores order students along a variable, they are described as having 'ordinal' properties.

It is common to assume that test scores also have 'interval' properties: that is, that equal differences in scores represent equal differences in the variable being measured (eg, that the difference between scores of 25 and 30 on a reading comprehension test represents the same difference in reading ability as the difference between scores of 10 and 15). The attempt to attribute interval properties to scores is an attempt to treat them as though they were measures similar to measures of length in centimetres or measures of weight in kilograms. But scores are not counts of a unit of measurement, and so do not share the interval properties of measures.

Scores are counts of items answered correctly and so depend on the particulars of the items counted. A score of 16 out of 20 easy items does not have the same meaning as a score of 16 out of 20 hard items. In this sense, a score is like a count of objects. A count of 16 potatoes is not a 'measure' because it is not a count of *equal units*. Sixteen small potatoes do not represent the same amount of potato as 16 large potatoes. When we buy and sell potatoes, we use and count a unit (kilogram or pound) which maintains its meaning across potatoes of different sizes.

A second reason why ordinary test scores do not have the properties of measures is that they are bounded by upper and lower limits. It is not possible to score below zero or above the maximum possible score on a test. The effect of these so-called 'floor' and 'ceiling' effects is that equal differences in test scores

do not represent equal differences in the variable being measured. On a 30-item mathematics test, a difference of one score point at the extremes of the score range (eg, the difference between scores of 1 and 2, or between scores of 28 and 29) represents a larger difference in mathematics achievement than a difference of one score point near the middle of the score range (eg, the difference between scores of 14 and 15).

Although test scores do not have interval properties, it is (mistakenly) common to treat them as though they do. Interval properties are assumed whenever number-right scores are used in simple statistical procedures such as the calculation of means and standard deviations, or in more sophisticated statistical procedures such as regression analyses or analyses of variance. In these common procedures, users of test scores treat them as though they have the interval properties of inches, kilograms or hours.

## Objectivity

Every test constructor knows that, in themselves, individual test items are unimportant. No item is indispensable: items are constructed merely as opportunities to collect evidence about some variable of interest, and every test item could be replaced by another, similar item. More important than individual test items is the variable about which those items are intended to provide evidence.

A particular item developed as part of a calculus test, for example, is not in itself significant. Indeed, students may never again encounter and have to solve that particular item. The important question about a test item is not whether it is significant in its own right, but whether it is a useful vehicle for collecting evidence about the variable to be measured (in this case, calculus ability).

It is a fundamental intention of all measures that their meaning should relate to some general variable, and should not be limited to the specifics of the instrument used to obtain them.

Another way of saying this is that it should not matter to our conclusion about a student's ability in calculus which particular items the student is given to solve. When we construct a test it is our intention that the results will have a generality beyond the specifics of the test items. This intention is identical to our intention that measures of height should not depend on the details of the measuring instrument (eg, whether we use a steel rule, a wooden rule, a builder's tape measure, a tailor's tape, etc). It is a fundamental intention of all measures that their meaning should relate to some general variable such as height, temperature, manual dexterity or empathy, and should not be bound to the specifics of the instrument used to obtain them. (Just imagine the inconvenience of physical measures if every time they were reported they had to be accompanied by information about the particular instrument used to obtain them!)

The intention that measures of educational variables should have a general meaning independent of the instrument used to obtain them is especially important when there is a need to compare results on different tests. A teacher or school wishing to administer a test prior to a course of instruction (a pre-test) and then after a course of instruction (a post-test) to gauge the impact of the course, often will not wish to use the same test on both occasions. A medical school using an admissions test to select

applicants for entry often will wish to compare results obtained on different forms of the admissions test at different test sittings. Or a school system wishing to monitor standards over time or growth across the years of school will wish to compare results on tests used in different years or on tests of different difficulty designed for different grade levels (eg, third, fourth, and fifth-grade reading tests).

There are many situations in education in which we seek measures that are freed of the specifics of the instrument used to obtain them and so are comparable from one instrument to another.

It is also the intention when measuring educational variables that the resulting measures should not depend on the persons doing the measuring. This consideration is especially important when measures are based on judgements of student work or performance. To ensure the objectivity of measures based on judgements it is usual to provide judges with clear guidelines and training, to provide examples to illustrate rating points (eg, samples of student writing or videotapes of dance performances), to use multiple judges, procedures for identifying and dealing with discrepancies, and statistical adjustments for systematic differences in judge harshness/leniency.

Although it is clearly the intention that educational measures should have a meaning freed of the specifics of particular tests, ordinary test scores (eg, number of items answered correctly) are completely test-bound. A score of 29 on a particular test does not have a meaning similar to a measure of 29 centimetres or 29 kilograms. To make any sense of a score of 29 it is necessary to know the total number of test items: 29 out of 30 items? 29 out of 40? 29 out of 100? Even knowing that a student scored 29 out of 40 is not very helpful. Success on 29 easy items does not represent the same ability as success on 29 difficult items. To understand completely the meaning of a score of 29 out of 40 it would be necessary to consider each of the 40 items attempted.

A longstanding dilemma in educational testing has been that, while particular test items are never of interest in themselves, but are intended only as indicators of the variable of interest, the meaning of number-right scores is always bound to some particular set of items. Just as we intend the measure of a student's writing ability to be independent of the judges who happen to assess that student's writing, so we seek measures of variables such as numerical reasoning which are neutral with respect to, and transcend, the particular items that happen to be included in a test. It is this dilemma that modern measurement theory (described in the next section) resolves.

## In Summary

In education, we seek measures on a wide variety of variables. Reliable measures of educational variables are essential to successfully evaluating the effectiveness of educational programs, monitoring educational standards over time, comparing achievement levels in different education systems, investigating relationships and influences on educational achievement, allocating scholarships and places in educational courses, measuring individual growth over time, and making decisions about the stage an individual has reached in his or her learning. Educational measurement always begins with the intention to estimate students' standings on some variable of interest.

In education, we assume that test and questionnaire scores have interval level properties whenever we calculate simple statistics such as means and standard deviations and use more sophisticated procedures such as regression analysis. However, ordinary test scores, because they are not counts of a unit of fixed amount, do not have interval properties. Equal differences in number-right scores do not, in general, represent equal differences on the variable of interest.

In education, we also intend our measures to have a generality that extends beyond the specifics of a set of items, and beyond the particular persons involved in the measuring process. Test items are not in themselves important: they are simply convenient and interchangeable opportunities to collect evidence about the variable a test is designed to measure. However, the meaning of number-right scores is bound to particular sets of items. Every test has its unique set of number-right scores – equivalent to every measuring stick being calibrated in its own unit of length.

The following article describes a measurement model that can be used to

- establish the extent to which a set of items work together to provide measures of just *one* variable;
- define a *unit* of measurement for the construction of interval-level measures of educational variables; and
- construct numerical measures which have a meaning *independent of* the particular set of items used.

# Falling Short of Measurement

## Norm-referenced testing

Ordinary test scores (counts of items answered correctly) do not share the properties of measures such as lengths in centimetres or weights in kilograms. A test score, such as 28 out of 40, does not have a general meaning like 28 centimetres because it is bound to a particular set of test items.

In an attempt to give test scores meanings beyond the specifics of a particular instrument, a common practice is to refer test scores to defined populations of students. For example, if 65 per cent of Scottish fourth-grade students scored below 28 on a particular test, then the score of 28 on that test would be re-expressed as the 65th percentile for Scottish fourth-graders. This method of interpreting test scores, while useful, does not provide a measurement scale with interval properties. It is equivalent to marking out a stick for measuring children's heights not in constant units such as inches or centimetres, but in percentiles. The 60th percentile would be the mark separating the shortest 60% of students from the tallest 40%; the 90th percentile would separate the shortest 90% of students from the tallest 10%; and so on. Of course, the 60th percentile for Year 4 students would be at a different position on the measuring stick from the 60th percentile for Year 3 students.

It is always possible to locate percentiles on a measuring stick calibrated in inches or centimetres. But percentiles are not a substitute for a well-defined unit and so do not provide a basis for measuring (counting units).

A more detailed discussion of the norm-referenced interpretation of scores can be found on page 38.

## Criterion-referenced testing

A second attempt to address the limitations of test-bound scores is to refer test results to narrowly defined domains of learning and to attempt to conclude in an absolute sense whether or not an individual has 'mastered' each domain. An example of such a domain might be 'subtracting two 2-digit numbers'. Has the student mastered the subtraction of 2-digit numbers or not? In a criterion-referenced test, a set of items is written to address each domain, and a student is considered to have 'mastered' the domain if they answer 80% of items correctly.

This method of interpreting test scores, while superficially attractive, has a number of shortcomings in practice. Even when domains are defined as narrowly as 'subtracting two 2-digit numbers', a student's success rate depends on the particular items administered. Subtraction items written vertically are easier than items written horizontally. Subtraction items requiring regrouping are harder than items not requiring regrouping. Subtraction items involving zeros are notoriously difficult. Whether a student answers 80% of items correctly depends on the particular items administered. In an attempt to remove ambiguities of this kind, a common approach is to define achievement domains more precisely (eg, subtracts two 2-digit numbers when written vertically and when regrouping is not required). But the consequence of this approach has been to fragment and atomise school curricula into increasingly long lists of increasingly trivial skills.

In practice, a criterion-referenced test is like a measuring stick with only one mark (the criterion). Students either meet the criterion or they do not. These domain-specific measuring sticks with their yes/no results provide a limited basis for monitoring individual growth in meaningful areas of learning over time.

# a model for measuring

The preceding article identified several reasons why ordinary test scores (counts of items answered correctly) do not have the properties of 'measures' such as lengths in centimetres or temperatures in degrees Celsius:

- Although the intention in most test development is to produce a single score for each student (in other words, to construct a single dimension along which students can be ordered from lowest to highest), much test development is not accompanied by an explicit check on the validity of summarising item responses in a single measure.
- Although the use of test scores in most statistical analyses assumes they have interval level properties, because test scores are not counts of a unit of measurement, ordinary test scores are not on an interval scale.
- Although our interest in educational testing is always in some underlying variable – and not in a specific set of items – ordinary test scores (eg, 28 out of 40) are always bound to a particular test, and so do not have instrument-neutral meanings (like 28 centimetres or 28 kilograms).

In short, ordinary 'number-right' scores do not have the properties of measures.

This article describes a method for constructing educational *measures*. These measures – when they can be constructed – share the properties described on pages 6 to 12. In other words, they are:

- estimates of locations on a single variable (unidimensional)
- expressed in a constant unit of measurement (interval-level)
- freed of the particulars of the instrument used (objective)

Measures with these properties do not come easily. The method described here requires data (observations) satisfying a demanding set of requirements. Although educational measures can be constructed from responses to test items, not every set of test items meets these requirements and is capable of yielding unidimensional, interval-level, objective measures.

How does the difficulty of the task (δ) compare with the person's ability (β)?

## One Variable

The model for measuring described here begins with the intention to focus on just one aspect of variability (ie, one variable) and to estimate individuals' locations on that one variable.

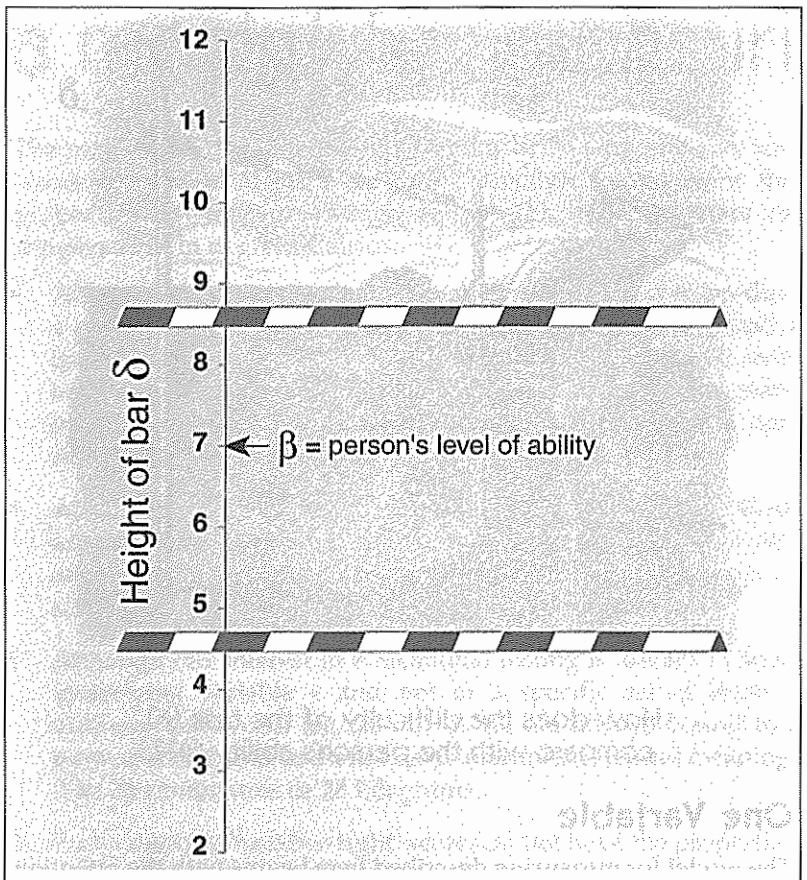
Suppose, for example, that the variable of interest is 'high-jumping' ability. We might hypothesise that individuals differ in their high-jumping abilities and that it is possible to obtain useful estimates of high-jumping ability by observing performances on some relevant 'high-jumping' tasks. Whether this idea is supported in practice will depend on the extent to which performances on the tasks we use are consistent with the proposition that individuals differ along a single dimension of high-jumping ability.

The notion of a high-jumping variable is represented in the picture on the next page. In this picture, high-jumping ability is imagined to increase up the page. The height of the bar determines the task difficulty, represented by the Greek letter delta (δ). Along this continuum of increasing difficulty, high-jumping abilities (β) of individuals also might be mapped. One individual's imagined ability $\beta = 7$ is marked.

Several observations can be made about this picture.

First, the high-jumping abilities (β) of individuals and the difficulties (δ) of high-jumping tasks can be conceptualised as positions along the *same* continuum. Easier tasks and individuals with lower abilities will be located towards the bottom of the continuum; harder tasks and individuals with higher abilities will be located towards the top.

Second, the high-jumping ability of this individual has been labelled with the Greek letter β (beta) to reflect the fact that this person's ability can never be known exactly – it can only be




A scale of increasing difficulty showing one person's level of ability (β)

imagined and then estimated from observations of the person's performances. The more high-jumping tasks the person attempts, the more information we will have, and the more confidence we will have in our estimate.

Third, this high-jumping variable has been marked out (optimistically) in what appear to be equal units. To develop *measures* of high-jumping ability, we require a constant unit of high-jumping ability.

## Planning Observations

To measure individuals on a variable, it is necessary to assemble evidence relevant to that variable. In the case of high-jumping ability, evidence in the form of observations of success or failure on a number of high-jumping tasks is likely to provide an appropriate basis for estimating individuals' abilities. For other variables, the most appropriate evidence might be collected using paper and pen tasks, or by judging portfolios of work, completed projects, or products such as items of technology or works of art.

To estimate a person's location on a variable it usually is not sufficient to observe the person's performance on, or response to, just one task. Success or failure on a single high-jumping task or a single question on a reading test provides very limited information about an individual's ability. Reliable measures require multiple observations.

Measurement requires observation under controlled conditions.

Measurement also requires observations under controlled conditions. The idea that individuals differ in high-jumping ability may originally have been developed from casual observations of people leaping over logs, rocks, hedges and fences. But to compare (and measure) high-jumping ability, we would not ask some individuals to jump a fence, others a hedge, and still others a rope. Rather, we would standardise the conditions of observation to minimise the influence of factors irrelevant to the variable of interest. The same is true of all measurement. For example, when we measure the heights of children, we measure them in a controlled and artificial situation – shoes off, chin up, and back to a wall.

## Records of Observations

Once a decision has been made about the assessment method to be used, a decision is required about the observations or judgements to be recorded. Here there are several possibilities. One possibility is to record *ratings* of individuals' performances or work. Judges' ratings commonly are used in the assessment of performances in areas such as gymnastics, public speaking, diving and instrumental music, in the assessment of student writing, and in assessing products of student work in technology and art. A second possibility is to use a system of *partial credit scoring* to identify students who give partially correct answers or who are partially successful in solving a problem. A third possibility is to use *dichotomous scoring* to record success or failure on a task. For example, students' responses to test questions often are recorded as either right or wrong. Individuals' attempts to clear a high-jump bar also are recorded dichotomously (cleared/missed).

It is usual to tabulate records of observations. The following table shows how a set of high-jump records might be tabulated for $_N$ persons. Each person's results are recorded in one row of the table. Individuals are assumed to have different abilities ($\beta_1, \beta_2, \beta_3, \ldots, \beta_N$). Each column corresponds to a particular height of the bar (with difficulty $\delta$) and the outcome of each attempt is recorded as either 1 (success) or 0 (failure).

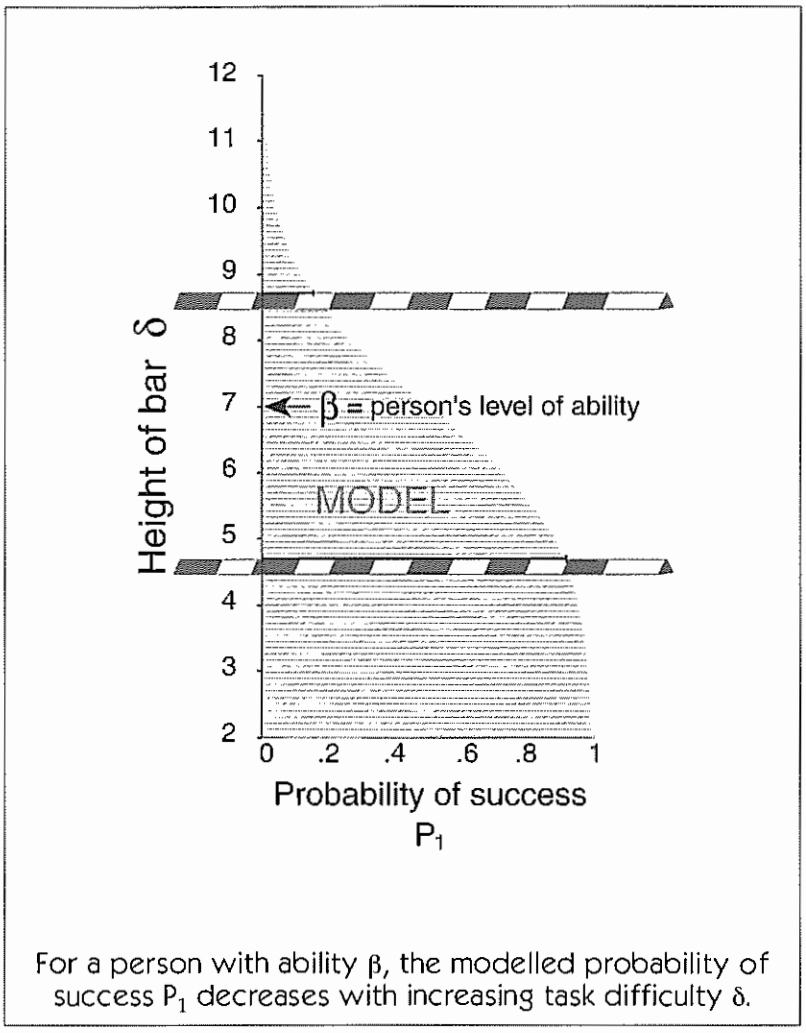
**Tasks**

| | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\delta_4$ | $\delta_5$ | ... | $\delta_L$ |
|---|---|---|---|---|---|---|---|
| $\beta_1$ | 1 | 1 | 1 | 1 | 0 | | 0 |
| $\beta_2$ | 1 | 1 | 0 | 1 | 0 | | 0 |
| $\beta_3$ | 1 | 1 | 1 | 1 | 0 | | 0 |
| $\beta_4$ | 1 | 1 | 0 | 0 | 0 | | 0 |
| $\beta_5$ | 1 | 0 | 1 | 0 | 1 | | 0 |
| ... | | | | | | | |
| $\beta_N$ | 1 | 1 | 0 | 1 | 1 | | 0 |

## A Measurement Model

The measurement model developed by Danish mathematician Georg Rasch provides a basis for estimating a person's ability $\beta$ from that person's row of recorded performances. The model proposes a mathematical relationship between a person's ability $\beta$, the difficulty $\delta$ of the task being attempted, and the probability $P_1$ of the person succeeding on that task.

This mathematical relationship is shown in the following picture.



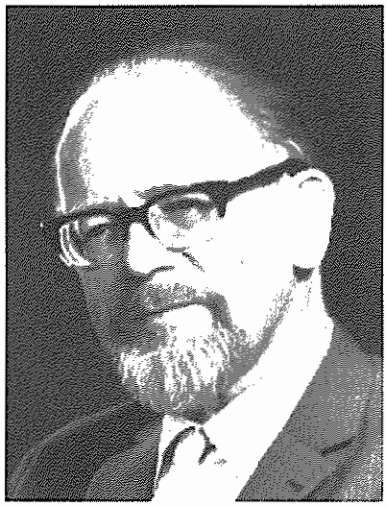

For a person with ability $\beta$, the modelled probability of success $P_1$ decreases with increasing task difficulty $\delta$.

This picture shows how, in the Rasch model, a person's probability of success $P_1$ decreases with increasing task difficulty. The relationship is shown here for a person of ability $\beta$=7. The more difficult the task (ie, the higher the bar), the lower the person's modelled probability of success.

When the result of a person's attempt at a task is scored 1 for success or 0 for failure, the person's probability of succeeding $P_1$ and probability of failing $P_0$ sum to one ($P_1+P_0=1$).

Expressed mathematically, the Rasch model gives the probability $P_1$ of a person with ability $\beta$ succeeding on a task of difficulty $\delta$ as:

$$P_1 = \exp(\beta-\delta) \ / \ 1+\exp(\beta-\delta)$$



Georg Rasch (1901—1980)

| $\delta$ | $P_1$ |
|---|---|
| 12.0 | .007 |
| 11.8 | .008 |
| 11.6 | .010 |
| 11.4 | .012 |
| 11.2 | .015 |
| 11.0 | .018 |
| 10.8 | .022 |
| 10.6 | .027 |
| 10.4 | .032 |
| 10.2 | .039 |
| 10.0 | .047 |
| 9.8 | .057 |
| 9.6 | .069 |
| 9.4 | .083 |
| 9.2 | .100 |
| 9.0 | .119 |
| 8.8 | .142 |
| 8.6 | .168 |
| 8.4 | .198 |
| 8.2 | .231 |
| 8.0 | .269 |
| 7.8 | .310 |
| 7.6 | .354 |
| 7.4 | .401 |
| 7.2 | .450 |
| 7.0 | .500 |
| 6.8 | .550 |
| 6.6 | .599 |
| 6.4 | .646 |
| 6.2 | .690 |
| 6.0 | .731 |
| 5.8 | .769 |
| 5.6 | .802 |
| 5.4 | .832 |
| 5.2 | .858 |
| 5.0 | .881 |
| 4.8 | .900 |
| 4.6 | .917 |
| 4.4 | .931 |
| 4.2 | .943 |
| 4.0 | .953 |
| 3.8 | .961 |
| 3.6 | .968 |
| 3.4 | .973 |
| 3.2 | .978 |
| 3.0 | .982 |
| 2.8 | .985 |
| 2.6 | .988 |
| 2.4 | .990 |
| 2.2 | .992 |
| 2.0 | .993 |

Notice that the probability of success $P_1$ depends on $\beta$-$\delta$ (in other words, on how far the bar is from the person's level of ability). When the bar is set at the person's ability, $\beta - \delta = 0$, and

$$P_1 = \exp(0) \; / \; 1+\exp(0) = 0.5$$

## A Unit of Measurement

The Rasch model can be rearranged as:

$$\beta - \delta = \ln(P_1 / P_0)$$

where the unit in which $\beta$ and $\delta$ are expressed is called a 'logit'. When the Rasch model is used to construct measures of ability, the resulting measurement variable is calibrated in logits.

The table on this page shows the modelled probability $P_1$ of a person with ability $\beta$=7 succeeding on tasks with difficulties $\delta$ ranging from 2 to 12 logits.

## The Key to Objectivity

A fundamental intention in all measurement is that measures of a variable should be independent of the details of the particular instrument used to obtain them. In educational measurement, our interest always is in the variable (ie, construct) to be measured, and not in any particular item or set of items. Every test is simply a convenient sample of many possible items that could be used in the collection of evidence about that variable.

At the most elementary level this intention means that, if we were to consider two persons A and B with assumed abilities $\beta_A$ and $\beta_B$ on the variable of interest, then our estimate of the difference $\beta_A$-$\beta_B$ between these two persons should not depend on which particular test items we happened to use to estimate this difference. If on one set of items person A was estimated to be, say, 1 logit more able than person B, then on any other set of items measuring that variable, person A should be estimated to be 1 logit more able than person B (within the limits of measurement error).

In our high-jumping analogy, we would hope that our estimate of the relative high-jumping abilities of persons A and B had some generalisable meaning – that its meaning was not limited to the few observations we had made or by the particular heights at which we had set the bar. Only if our estimates of the relative abilities of individuals are generalisable to tasks beyond those used to obtain them do we have any hope of constructing 'measures' of variables.

When two persons A and B attempt the same dichotomously scored task, there are four possible outcomes: both persons succeed (✓✓); person A succeeds but B fails (✓x); person A fails but B succeeds (x✓); and both persons fail (xx). Only two of these outcomes (✓x and x✓) contain information about the *relative* abilities of persons A and B.

If the probability of person A succeeding on the task is $P_A$, and the probability of person B succeeding is $P_B$, then the probabilities of these four possible outcomes are given by the following joint probabilities (see note on page 22):

| | | | |
|---|---|---|---|
| both persons succeed | $P_{✓✓}$ | = | $P_A \times P_B$ |
| person A succeeds but B fails | $P_{✓x}$ | = | $P_A \times (1-P_B)$ |
| person A fails but B succeeds | $P_{x✓}$ | = | $(1-P_A) \times P_B$ |
| both persons fail | $P_{xx}$ | = | $(1-P_A) \times (1-P_B)$ |

The conditional probability of person A succeeding and B failing, given that one person succeeds and the other fails, is:

$$P_{✓x} / (P_{✓x} + P_{x✓}) = P_A \times (1-P_B) / [P_A \times (1-P_B) + (1-P_A) \times P_B]$$

And the conditional probability of person A failing and B succeeding, given that one person succeeds and the other fails, is:

$$P_{x✓} / (P_{✓x} + P_{x✓}) = (1-P_A) \times P_B / [P_A \times (1-P_B) + (1-P_A) \times P_B]$$

From these two equations it follows that:

$$P_{✓x} / P_{x✓} = \exp(\beta_A - \delta) / \exp(\beta_B - \delta) = \exp(\beta_A - \beta_B)$$

In other words,

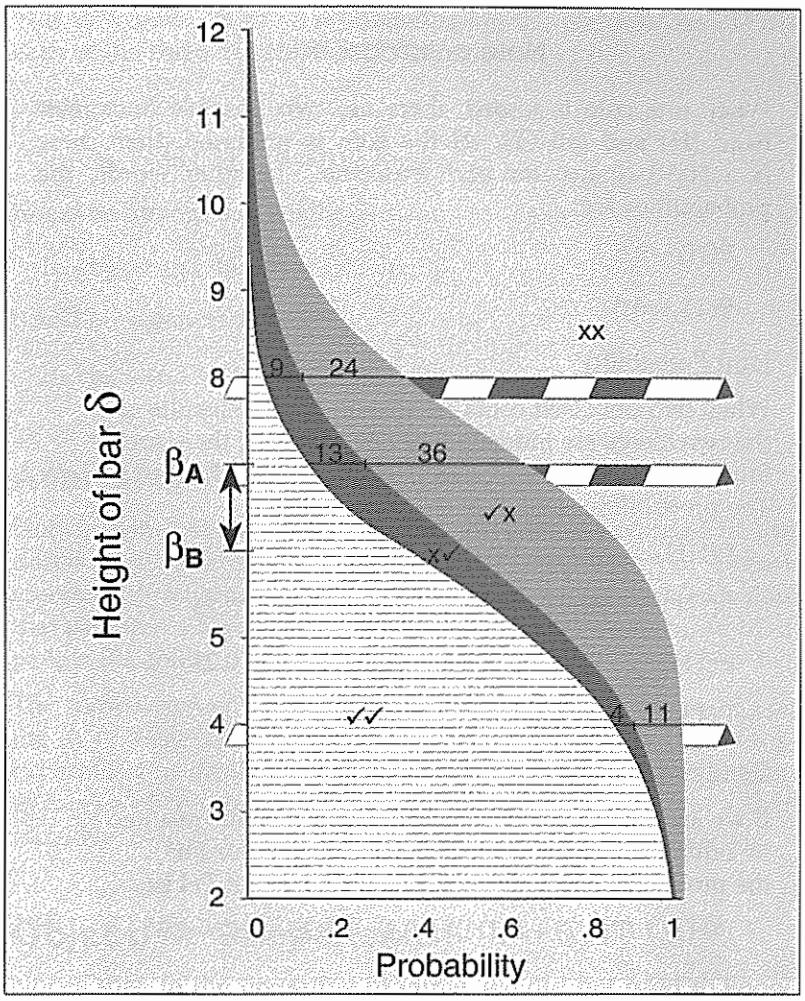
$$\beta_A - \beta_B = \ln(P_{✓x} / P_{x✓})$$

The implications of this feature of the Rasch model are shown in the picture on the page opposite.

In this picture, the high-jumping abilities ($\beta_A$=7 logits; $\beta_B$=6 logits) of two persons A and B are marked. The graph shows how the Rasch probabilities of persons A and B both succeeding (✓✓), A failing and B succeeding (x✓), A succeeding and B failing (✓x), and both failing (xx) vary with task difficulty $\delta$.

The important point in this picture is that the ratio of the width of the lighter grey region ($P_{✓x}$) to the width of the darker grey region ($P_{x✓}$) is *constant* for all values of $\delta$ (eg, .24/.09 = .36/.13 = .11/.04).

The significance of this feature of the Rasch model is that it is not necessary to know or to estimate the height of the bar (difficulty of the task) to estimate the relative abilities of persons A and B. If these two persons were given multiple attempts at clearing the bar at *any* particular height, an estimate of their relative abilities




could be obtained from the number of times A succeeded and B failed ($N_{✓x}$) and the number of times A failed and B succeeded ($N_{x✓}$). All that is required is to calculate the proportions

$$p_{✓x} = N_{✓x} / (N_{✓x}+N_{x✓})$$
$$p_{x✓} = N_{x✓} / (N_{✓x}+N_{x✓})$$

and to substitute into:

$$b_A - b_B = \ln (p_{✓x} / p_{x✓}) = \ln (N_{✓x} / N_{x✓})$$

Where $b_A - b_B$ is an estimate of the difference $\beta_A - \beta_B$. In other words, if a set of high-jumping data conform to the Rasch model, then the difference ($\beta_A - \beta_B$) between persons A and B can be estimated (in logits) by setting the bar at *any* height and simply counting the results ✓x and x✓.

### Note on probabilities

When one coin is tossed, there are two possible outcomes: head (H) and tail (T). If the coin is unbiased, then there is a 50:50 chance for each outcome. In other words, the probabilities are:

$P(H) = 0.5$

$P(T) = 0.5$

Joint Probability

When two unbiased coins are tossed, and the results of the two tosses are independent of each other, there are four equally likely outcomes: HH, HT, TH, and TT. The probabilities are:

$P(HH) = P(H) \times P(H) = 0.5 \times 0.5 = 0.25$

$P(HT) = P(H) \times P(T) = 0.5 \times 0.5 = 0.25$

$P(TH) = P(T) \times P(H) = 0.5 \times 0.5 = 0.25$

$P(TT) = P(T) \times P(T) = 0.5 \times 0.5 = 0.25$

Conditional Probability

If we are told that a particular toss of two coins resulted in 'odds' (HT or TH) rather than 'evens' (TT or HH), the probability that the result was HT and not TH is:

$P(HT) / (P(HT) + P(TH)) = 0.25 / (0.25+0.25) = 0.5$

## An Illustration

To illustrate this fundamental feature of the Rasch model, we now consider the results of a hypothetical high-jumping exercise.

Suppose that, to estimate the relative high-jumping abilities of persons A and B, we set the bar at three different heights (i, ii and iii) and recorded the two jumpers' results on each of 100 attempts at each height:

| Height of bar | A and B both clear ✓✓ | A clears B misses ✓x | B clears A misses x✓ | A and B both miss xx | Total attempts |
|---|---|---|---|---|---|
| iii (hard) | 1 | 11 | 4 | 84 | 100 |
| ii | 14 | 36 | 13 | 37 | 100 |
| i (easy) | 64 | 24 | 9 | 3 | 100 |

Only the shaded part of this table contains information about the relative abilities of persons A and B. We can now use the equation at the bottom of page 21 to estimate the difference between the high-jumping abilities of persons A and B:

$b_A - b_B = \ln(N_{✓x} / N_{x✓})$

This difference could be estimated using the jumpers' results on each of the three heights separately:

based on attempts at height i

$b_A - b_B = \ln(24 / 9) = 0.98$ logits

based on attempts at height ii

$b_A - b_B = \ln(36/13) = 1.02$ logits

based on attempts at height iii

$b_A - b_B = \ln(11 / 4) = 1.01$ logits

## Analysing Fit

Notice that, because the 'results' in the table on page 22 fit the Rasch model extremely well, the three estimates are almost identical. Attempts at heights i, ii and iii all lead to the conclusion that person A's high-jumping ability is about *one logit* greater than person B's ability.

The comparison of these three estimates provides a simple test of the fit of these data to the Rasch model. If, in a real high-jump experiment of this kind, the three heights did not lead to similar estimates of the difference between persons A and B, then that would be evidence that the data did not conform to the model.

## Objective Comparisons

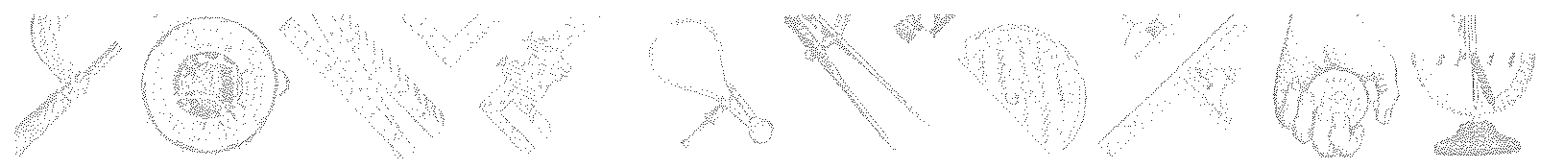
Notice that, in the preceding example, we did not need to know the heights i, ii and iii of the bar (in inches, centimetres or logits) to estimate the relative abilities of persons A and B. When data fit the Rasch model, it is possible to compare abilities without knowing, or even having to estimate, the difficulties of the tasks. This is a unique feature of the Rasch model.

To further illustrate the point that Rasch ability estimates are independent of the difficulties of the tasks, notice that we could have estimated the relative abilities of persons A and B from the results of any 200 attempts:

| | | | |
|---|---|---|---|
| i and ii | $b_A - b_B$ | $= \ln((11+36)/(4+13))$ | $= \ln(47/17)$<br>$= 1.02$ logits |
| ii and iii | $b_A - b_B$ | $= \ln((36+24)/(13+9))$ | $= \ln(60/22)$<br>$= 1.00$ logits |
| i and iii | $b_A - b_B$ | $= \ln((11+24)/(4+9))$ | $= \ln(35/13)$<br>$= 0.99$ logits |

or – the best estimate of all – from the results of all 300 attempts:

$b_A - b_B = \ln((11+36+24)/(4+13+9))$
$= \ln(71 / 26) = 1.00$ logits



Because it is possible to simply sum down the middle columns of the table on page 22 in this way, without regard to the difficulties of the tasks, there is no reason to require persons A and B to make more than one attempt at any given height. The two jumpers could make one attempt each at, say, L different heights, with the outcomes recorded in a table such as this:

| Height number | A and B both clear ✓✓ | A clears B misses ✓x | B clears A misses x✓ | A and B both miss xx | Total attempts |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 |
| 2 | 0 | 1 | 0 | 0 | 1 |
| 3 | 1 | 0 | 0 | 0 | 1 |
| ... | | | | | 1 |
| L | 0 | 0 | 1 | 0 | 1 |

Once again, the relative abilities of persons A and B could be estimated by summing down the middle columns of the table to obtain $N_{✓x}$ and $N_{x✓}$ and then substituting into:

$$b_A - b_B = \ln (N_{✓x}/N_{x✓})$$

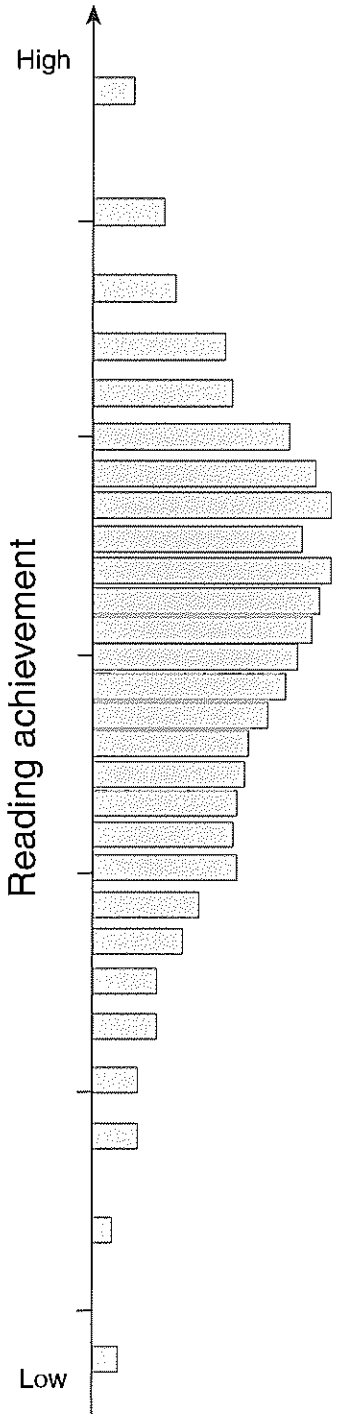
If the performances of persons A and B were consistent with their performances in the table on page 22, then the totals of the two middle columns would have a ratio of about 2.7:1, leading to an estimated difference of about 1.0 logits, regardless of the heights they attempted.

In the above case, the fit of the data to the model might be tested by comparing the estimate $b_A - b_B$ based on attempts at the first L/2 heights with the estimate based on attempts at the second L/2 heights; or by comparing the estimate based on all the even-numbered heights with the estimate based on all the odd-numbered heights. These four estimates will be very similar when the recorded observations fit the Rasch model.

## Application to Test Data

The procedure just applied to high-jumping data also can be applied to a set of test data. In a test, each respondent has only one attempt at each item. If that attempt is recorded as either right (✓) or wrong (x), then the test performances of two respondents A and B on a test of length L could be summarised in a table as follows:

| Item number | A and B both right ✓✓ | A right B wrong ✓x | B right A wrong x✓ | A and B both wrong xx | Total attempts |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 |
| 2 | 0 | 1 | 0 | 0 | 1 |
| 3 | 1 | 0 | 0 | 0 | 1 |
| ... | | | | | 1 |
| L | 0 | 0 | 1 | 0 | 1 |




Distribution of estimated student locations on a reading achievement variable

The relative abilities of persons A and B are estimated by summing down the middle two columns of the table to obtain $N_{\checkmark x}$ and $N_{x\checkmark}$ and then substituting into:

$$b_A - b_B = \ln (N_{\checkmark x} / N_{x\checkmark})$$

The fit of these test data to the Rasch model could be tested by comparing estimates obtained on different subsets of items (eg, even items, odd items; first half of test, second half of test).

## Estimating Locations on a Variable

If all possible pairs of students taking a test are considered in this way, and all these 'pairwise' estimates are brought together, then it is possible to estimate the locations of all students taking the test along the same variable. This procedure is known as the 'pairwise' method of estimating students' abilities on a variable.[1]

The result of applying the pairwise estimation procedure to a set of test data is an estimate of each student's location on the variable the test is designed to measure. The diagram on this page shows a distribution of Year 3 students' estimated locations on a variable of increasing reading achievement. These estimated locations are mapped on an interval-level scale marked out in logits.

## In Summary

The Rasch model described in this article specifies the requirements a set of test data must meet if they are to provide *measures* which are: (i) estimates of individuals' locations on a single variable/dimension; (ii) expressed in a constant unit of measurement; and (iii) freed of the particulars of the instrument used to obtain them.

Measures with these properties do not come easily. Although they can be constructed from responses to test items, not every set of test items meets these requirements and is capable of yielding unidimensional, interval-level, objective measures.

The key to objective measurement resides in the fact that, when two persons A and B attempt an item, under the Rasch model, the ratio $P_{\checkmark x} / P_{x\checkmark}$ is governed only by the relative abilities of the two persons:

$$\beta_A - \beta_B = \ln(P_{\checkmark x} / P_{x\checkmark})$$

(where $P_{\checkmark x}$ is the modelled probability of person A succeeding and B failing the item, and $P_{x\checkmark}$ is the probability of A failing and B succeeding). It is this feature of the model that makes possible measures which are 'freed' of the particulars of the items used to obtain them.

When the Rasch model is applied, it provides a measure for each student on a continuum marked out in equal intervals called 'logits'.

[1] Choppin, B. (1976). Recent developments in item banking. In DN de Gruitjer & LJ Vanderkamp (eds). *Advances in Psychological and Educational Measurement*. London: Wiley.

# mapping variables

A fundamental characteristic of 'measures' is that they indicate positions on general variables. In other words, they have meanings that are not limited to, and can be generalised beyond, the specific instruments used to obtain them. For example, measures of length in centimetres indicate positions on the general variable 'length' and have meanings that do not depend on the details of the instrument used (eg, wooden rule, steel tape measure, callipers, dressmakers' tape).

Educational measures also are intended to indicate positions on general variables. For example, measures of reading ability are intended to indicate positions on the general variable 'reading ability' and to have meanings that are not limited to the particular passages of text or particular test questions used to obtain them. Test developers know that individual test questions are never of significance in themselves: they are simply opportunities to collect samples of behaviour for the purposes of estimating positions on the general variable of interest.

When it comes to *interpreting* educational measures, it is important to look beyond the specifics of the instrument to the generalities of the underlying measurement variable. It is to this topic that we now turn.

## Marking Out a Variable

In our discussion of the measurement of high-jumping ability we noted that the difficulties of high-jumping tasks and the abilities of individuals could be conceptualised as positions on the same variable. In real high-jumping events, the difficulty of a task is determined by the height of the bar from the ground. When the bar is set at increasing heights, these increasingly difficult tasks define increasing levels of high-jumping ability.

But an alternative to measuring the height of the bar from the ground would be to estimate the difficulty of each high-jumping task from records of jumpers' success rates on that task. If a group of jumpers attempted the same set of tasks, then the height cleared by the greatest number in the group would be estimated to be the easiest, and the height cleared by the smallest number would be estimated to be the hardest. Using the measurement model on page 18, the difficulty of each task could be estimated (in logits) from the available records of jumpers' performances.

This process of estimating the difficulties of a set of tasks is known as 'calibration'. To illustrate the calibration process it is convenient to begin by considering one individual's attempts at two high-jumping tasks Y and Z with difficulties $\delta_Y$ and $\delta_Z$. If the person has one attempt at each height, then there are four possible outcomes: the person succeeds on both (✓✓); succeeds on Y but fails Z (✓x); succeeds on Z but fails Y (x✓); and fails both (xx).

Only two of these four possible outcomes (✓x and x✓) are useful in estimating the relative difficulties of the two tasks.

Following steps parallel to those outlined on page 20 – which include calculating the conditional probability of the person succeeding on each task, given that they succeed on one but fail the other – the distance between tasks Y and Z on the variable is:

$$\delta_Z - \delta_Y = \ln (P_{\checkmark x}/P_{x\checkmark})$$

If this person attempts tasks Y and Z on a number of occasions, and on each occasion a record is kept of whether the outcome is xx, ✓x, x✓ or ✓✓, then the distance between tasks Y and Z can be estimated as:

$$d_Z - d_Y = \ln (n_{\checkmark x}/n_{x\checkmark})$$

where $d_Z - d_Y$ is an estimate of $\delta_Z - \delta_Y$, $n_{\checkmark x}$ is the number of times the person succeeds on Y but fails Z, and $n_{x\checkmark}$ is the number of times the person succeeds on Z but fails Y.

The important observation here is that this estimate does not depend on the person's ability. The distance between tasks Y and Z can be estimated by counting the outcomes ✓x and x✓ for *any* person. And, when data conform to the model, the estimates obtained in this way from the performances of different individuals are statistically equivalent.

From this observation it follows that, to estimate the distance between tasks Y and Z, it is not necessary to ask individuals to make more than one attempt at each task. For any group of persons, all that is required is that a record be kept of the number of ✓x and x✓ outcomes *for the group*. The distance can then be estimated as:

$$d_Z - d_Y = \ln (N_{\checkmark x}/N_{x\checkmark})$$

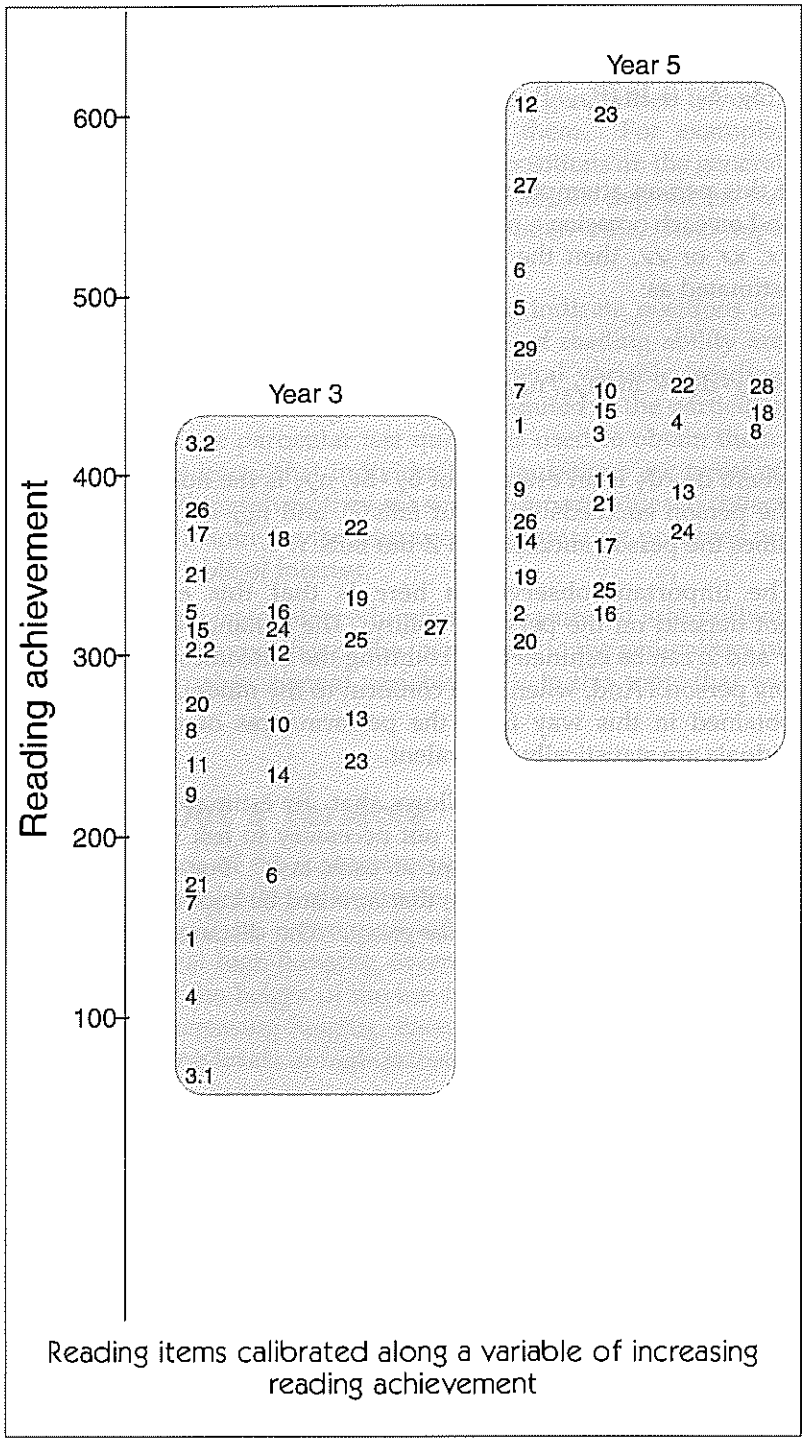
where $N_{\checkmark x}$ is the number of persons succeeding on Y but failing Z, and $N_{x\checkmark}$ is the number of persons succeeding on Z but failing Y.

This process can be repeated for all possible pairs of tasks, and the estimated distances between tasks brought together to calibrate all tasks along the same variable.

When this procedure is applied to records of students' performances on a set of test items, an estimate (in logits) is obtained of each item's difficulty, allowing all items to be calibrated along the variable on which students are measured.

The diagram on the next page shows two sets of items calibrated along a variable of increasing reading achievement. The items shown here were administered in two tests: a test for third grade students and a test for fifth-grade students. The easiest item (numbered 3.1 on the Year 3 test) is at the bottom of the diagram; the hardest (numbered 12 on the Year 5 test) is at the top. From this diagram it is clear that the Year 5 test was generally more

difficult than the Year 3 test, although there are many items on both tests in the range 300 to 400. (The numbers on the vertical scale are multiple logits.)




Reading items calibrated along a variable of increasing reading achievement

When a number of test items are calibrated along a variable in this way, the locations of individual items provide insights into the underlying variable. Each item is an example of the variable in the region in which it is calibrated. For example, item 3.1 above is a relatively easy reading item requiring a relatively low level of reading ability. This item requires Year 3 children to look at the cover of an age-appropriate storybook and to identify key elements of the story from the book title and illustration. Children with very low reading abilities (below about 100 on this

'John Ogilby's publication in 1675 of maps of the main routes out of London was the first of its kind anywhere in Europe. The roads were illustrated in strips going parallel up the page, regardless of the direction they took on the ground. They were illuminated with drawings of the instruments used in their construction: road wheels to measure distance, quadrants and surveyors' chains. Features along the roads were also illustrated.'

John Ure

scale) will probably not be able to complete tasks of this kind. Item 3.1 is the only item on these two tests illustrating this level of early reading development.

At the other extreme, the most difficult reading items on these two tests are items 12, 23 and 27 on the Year 5 test. To answer these items correctly students have to interpret the expression 'last but not least', infer meaning from figurative language, and demonstrate an understanding of the connection between the content and form of a piece of text. These are examples of relatively high levels of reading skill in the region of 600 on this scale.

An analysis of what students have to do to provide the correct answer to each item on these two tests provides the beginnings of a 'map' of reading development across the third to fifth grades of school. The reading achievement map on page 30 summarises the skills assessed by most of the items on these two tests. A more detailed version of this map would include examples of the test items themselves to illustrate positions along the map. And a still richer understanding would be obtained by adding other calibrated items to this picture and investigating typical features and demands of items at various locations along the continuum.

## Analysing Stability

An important question when mapping variables in this way is whether the locations of items along the variable are stable across the students being measured. Individuals can be measured and compared meaningfully on a variable only if the variable itself is stable.

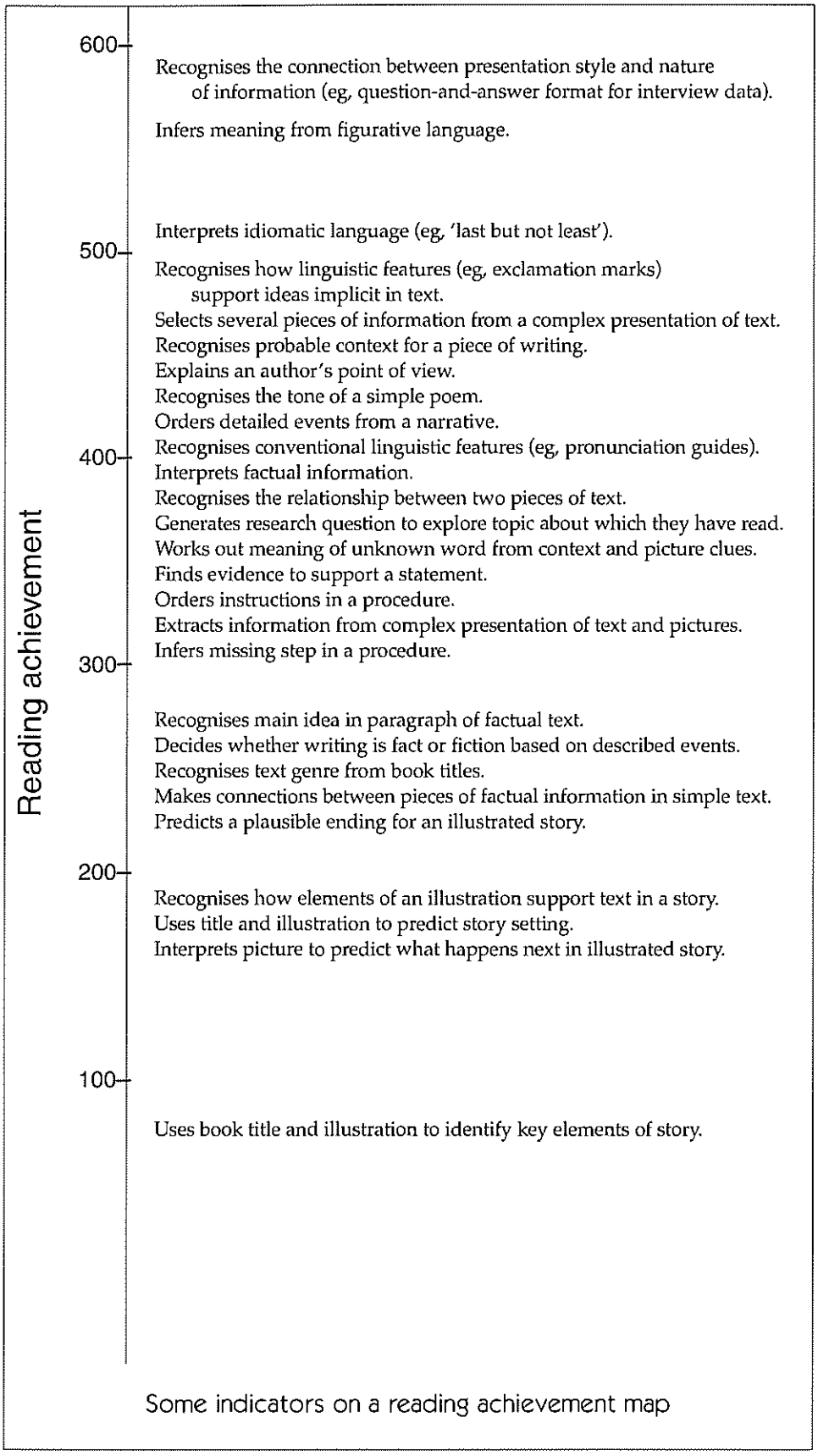
At the most elementary level we can ask whether the same estimated difference $d_Z - d_Y$ in the difficulties of two items Y and Z is obtained from the responses of different groups of students (eg, male, female, high-scoring, low-scoring, odd-numbered, even-numbered students). This test can be conducted by counting the students in each group with Y right and Z wrong ($N_{✓x}$) and the number of students with Y wrong and Z right ($N_{x✓}$), and then estimating the difference as:

$$d_Z - d_Y = \ln (N_{✓x} / N_{x✓})$$

When observations fit the Rasch model, this difference is the same (statistically equivalent) for different student subgroups. More generally, if an instrument is stable in its functioning across the students with whom it is to be used, then for *any* given item pair (i,j), statistically equivalent estimates of the difference $d_j - d_i$ will be obtained from the responses of different subgroups of students.

The statistical analysis of the stability of item difficulties across different student subgroups is known as 'differential item functioning' (*dif*) analysis.

The graph on page 31 is a pictorial display of the results of a *dif* analysis. This graph was constructed by calibrating the items on a statewide primary-school reading test on male and female




Some indicators on a reading achievement map

'If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement.'

(Thurstone, 1928, 547)[1]

pupils separately and then plotting these two sets of difficulty estimates against each other. The easiest item in this set (for both males and females) is at bottom left; the hardest is at top right. The unshaded band shows the region of statistical equivalence under the model (95% confidence region).

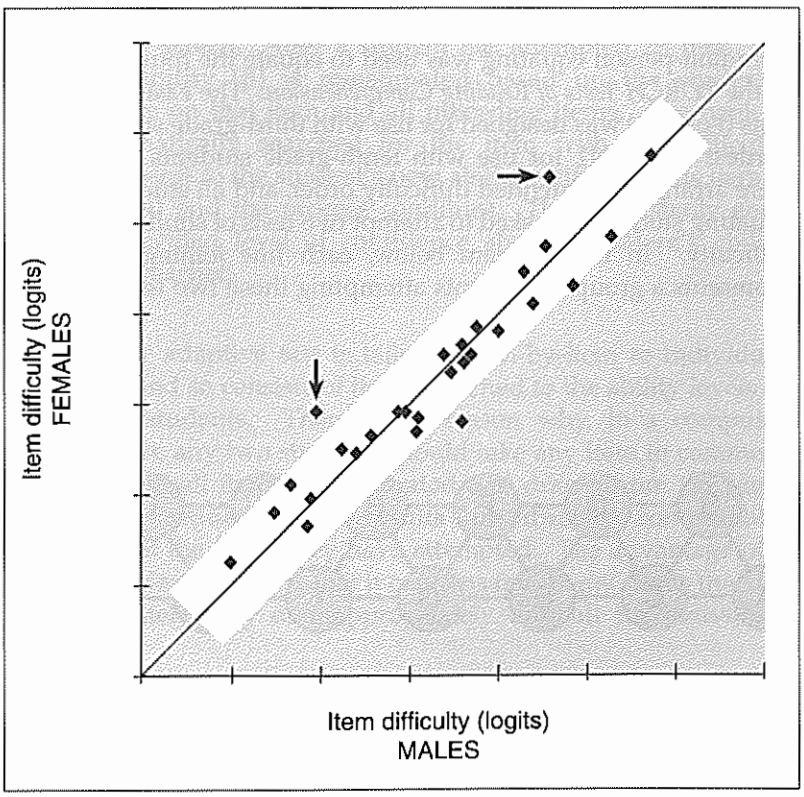
Two items (arrowed) are located above the band. Relative to the other items on this test, these two items are significantly more difficult for females than for males. A question that might be asked about these two items is whether their content places females at a special disadvantage. A third item is just below the band and, relative to the other items on this test, is more difficult for boys than for girls.

When used routinely in test development, *dif* analyses provide a basis for identifying items that may be biased against particular groups of students. Only if items retain their relative difficulties throughout the student population with which they are to be used (ie, are 'unbiased') do they provide the stability required of a measuring instrument.

## Item Banks

The map on page 30 shows reading items from two tests calibrated on a continuum of increasing reading achievement. Other reading items could be developed and calibrated along this continuum, provided that responses to those items also fit the Rasch model. In theory, there is no limit to the number of items that could be calibrated along a variable, and the larger the number of calibrated items, the richer the description and illustration of that variable.

A collection of calibrated test items is referred to here as an 'item



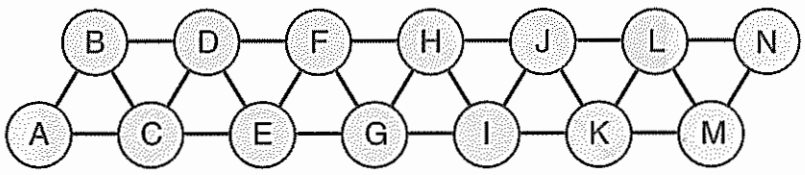
bank'. Some writers use the term 'item bank' to refer to any collection or pool of test questions. We follow the convention proposed by Bruce Choppin of reserving the term 'bank' for a set of items calibrated together on a common measurement variable. Only if items have been calibrated together along a common variable do they constitute a bank.

An item bank is constructed by jointly calibrating items from different tests or by undertaking special 'equating' studies in which students take items from more than one test. The items on pages 28 and 30 were calibrated along the same variable using the fact that some of the Year 3 items also were included in the Year 5 test. These common items provided the 'link' required for the joint calibration of the two tests. In the calibration process, the total set of Year 3 and Year 5 items was in effect treated as one large test in which only some items (the common items) were taken by all students.

Further items could be calibrated on this reading variable by embedding some of the items on page 28 into new tests as they were developed. The already calibrated bank items would provide the link required to bring new items on to the variable. (In practice, this could be done by independently estimating the difficulties of all items on a new test and then adjusting each of the item difficulties by the amount required to make the average difficulty of the common items the same as their average in the bank.) This process is known as 'common-item' equating.

An alternative procedure for calibrating a large number of items on the same variable is to ask groups of students to take more than one test. A group of students taking two tests provides the link required to calibrate those two tests on the same variable. In the calibration process, the two tests are in effect treated as one large test. This process is known as 'common-person' equating.

Common-person equating was used to equate the fourteen forms of the *TORCH Tests of Reading Comprehension*. The easiest of these tests (Form A) was designed for use with third-grade students; the hardest (Form N), for use with tenth-grade students. The tests were arranged in intended difficulty order, and all students in the equating study were asked to attempt two tests of similar intended difficulty. In the diagram below, each line joining two tests represents a group of students attempting those two tests.



The linking of the fourteen *Tests of Reading Comprehension* in this way allowed all *TORCH* items to be calibrated along a continuum of increasing reading ability. Teachers using *TORCH* choose a test appropriate to students' current reading abilities. Because all tests are calibrated along the same variable, performances on one test can be compared directly with performances on any other test, and reading growth can be monitored over time.

Item banks vary in size from several dozen items to many thousands of items. Once a bank has been constructed, it can be used as a source of calibrated items for the construction of new test forms. Any combination of calibrated items selected from an item bank is capable of providing student measures on the bank variable. When students' responses conform to the Rasch model, these measures are directly comparable with measures based on any other selection of bank items.

The advantages of an item bank include the fact that it is not necessary to administer exactly the same test items to all students. A set of relatively easy items can be selected and administered to students with relatively low levels of achievement, a set of more difficult items can be administered to more able students, and the results on the two tests can be compared directly. Student measures of this kind are 'objective' in the sense described on pages 10 and 11 – their meaning does not depend on knowledge of the particular items used to obtain them.

## Computer Adaptive Testing

When items are drawn from a calibrated-item bank, and students' responses conform to the Rasch model, it is possible to compare directly the performances of students taking different selections of test items. In a computer adaptive test, items are presented one at a time on a screen. After a student has attempted an item, the student's ability ($\beta$) is re-estimated based on the student's performance on that item and all preceding items. The bank is then automatically searched for the item with the difficulty estimate closest to the student's new ability estimate. This item is administered and the process continues. The test usually ends when a specified level of confidence about a student's ability estimate is reached.

A computer adaptive test is tailored item-by-item to individual test takers and so consists of items matched to the ability levels of individual students. There is no reason why, in a computer adaptive test, any two students should take any item in common. And, because all items are calibrated and drawn from the same item bank, students' test results are directly comparable, regardless of the items they have attempted. The advantage of a computer adaptive test is that it contains few, if any, items that are inappropriately easy or inappropriately difficult for individual students.

## In Summary

When items are calibrated along a variable, they begin to give meaning to that variable. They indicate typical observations at particular locations along the variable. When considered together, calibrated test or questionnaire items provide insights into the nature of typical progress or development. They form a 'map' against which student progress can be observed and monitored. The larger the number of items calibrated along a variable, the more richly the variable can be described and illustrated.

The key to the objective calibration of test or questionnaire items resides in the fact that, under the Rasch model, when an individual attempts two items Y and Z, the ratio $P_{\checkmark x} / P_{x\checkmark}$ is governed only by the relative difficulties of the two items:

$$\delta_Z - \delta_Y = \ln (P_{\checkmark x}/P_{x\checkmark})$$

(where $P_{\checkmark x}$ is the modelled probability of the individual succeeding on Y but failing Z, and $P_{x\checkmark}$ is the probability of the individual failing Y but succeeding on Z). When data conform to the Rasch model, it is possible to estimate the relative difficulties of any pair of items Y and Z from a simple record of student performances on those two items, regardless of the students involved.

In practice, it is essential that checks are made on the extent to which observations conform to the model. Only if item difficulty estimates are stable across the student population with which they are to be used can all students in that population be measured and compared on the same variable. Checks on differential item functioning indicate the extent to which a variable maintains its meaning across particular subgroups of the student population.

When a large number of items are calibrated on a variable, they constitute an item 'bank'. An advantage of an item bank is that it allows items to be selected and combined into different tests and students' performances on these different tests to be compared directly. A computer adaptive test draws on a bank of calibrated items to construct tests tailored to the item-by-item performances of individual test takers.

[1] Thurstone, LL (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.

# reporting measures

Educational measurement begins with the intention to estimate students' locations on some variable of interest. In education we are interested in many different aspects of student development, including reading ability, scientific literacy, respect for other cultures, mathematical competence, love of learning, logical reasoning, proficiency in the use of technology, and interpersonal skills. Every attempt to measure is an attempt to establish students' current levels of attainment in some aspect of their development.

Measuring instruments – tests and questionnaires – are designed to provide observations that can be used to estimate levels of attainment. But particular measuring instruments are never important in themselves: every test item can be replaced by one of many other equally appropriate items; every test can be replaced by some alternative selection of items. In education we intend our measures to have a generality that extends beyond the specific set of items used to obtain them. We are interested in a student's performance on a particular selection of items only to the extent that it indicates the student's standing on the underlying variable that the test is designed to measure. This intention is common to all measurement. For example, when we use a set of bathroom scales, we expect the particular bathroom scale we use to be irrelevant to the result, and the measure of our weight to be expressed in a metric that is not peculiar to that instrument.

If measures are to be compared meaningfully from one instrument to another (eg, from one reading test to another), and if they are to be used to measure change and to monitor growth, then they must be reported on measurement scales that are not tied to any one instrument.

## Interpreting Measures

In educational measurement, our primary interest when interpreting and reporting student attainment is usually in the knowledge, skills, understandings, attitudes or values that students have acquired. We also may be interested in comparing students' levels of attainment with the attainments of other students (eg, students of the same age or grade, students in other States or countries) or in knowing what progress students have made since some earlier occasion. But for many purposes – and particularly for the purposes of instruction – our primary interest is in knowing how students are progressing in relation to some continuum of developing knowledge, skills, understandings, attitudes or values.

To interpret educational measures in this way, it is first necessary to give substantive meaning to the variable being measured by mapping the kinds of observations typically made at varying locations along that variable. An example of such a mapping is shown on page 37. The numbers on the left of page 37 indicate increasing levels of reading ability as defined by the Lexile Framework for Reading[1]. The literature titles are examples of

books at different levels of reading difficulty. The easiest book shown here is *Ronald Morgan Goes to Bat* (200 lexiles); the most difficult is *Jonathan Livingston Seagull* (990 lexiles). A student with a reading level of, say, 880 should be able to read text at that level (eg, *The Red Pony*) with 75 per cent comprehension. On the right of the page are examples of texts at increasing levels along this continuum.

The progress map on page 37 allows students' reading abilities (measured on the Lexile scale) to be interpreted in terms of the kinds of texts they are likely to be able to read and understand, and suggests books that might be appropriate at particular levels of reading ability. (The appropriateness of a book for a particular student will depend, of course, not only on its difficulty level but also on its content and level of interest for the child.)

Every measurement variable in education can be mapped and illustrated with examples of the kinds of skills, responses or behaviours that characterise levels of development along that variable. Descriptions and illustrations attach substantive meaning to a variable and clarify the nature of growth in the area being measured. Progress maps of this kind are sometimes called 'described proficiency scales', and the process of interpreting students' levels of attainment with reference to such maps, 'standards referencing'.

In the construction of measurement variables it is common to define broad levels of attainment and to describe and illustrate typical observations within each level. The proficiency scale in Civics on page 40 was constructed from an analysis of US students' performances in the National Assessment of Educational Progress. The numerical scale on the left of the page is divided into four broad levels, and the kinds of knowledge and understanding typical of students at each level are described.[2] As in the Lexile example, these described levels provide a frame of reference for interpreting measures of attainment.

## Performance Standards

As well as providing frames of reference for reporting and describing students' current levels of attainment and charting progress over time, measurement variables of the kind illustrated on pages 37 and 40 also can be used in setting expectations or targets for student achievement. For example, the map on page 37 could be used to identify a level of reading ability that might reasonably be expected of all students by the end of fourth grade. The map of developing Civics knowledge on page 40 might be useful in thinking about the level of Civics knowledge to be set as a target for all eighth-grade students. Expectations or targets for student achievement also are called 'performance standards'.

A performance standard identifies the kinds of skills or understandings expected of students and, operationally, takes the form of a minimum score that must be achieved if a student is to be considered to have met the standard (eg, 270 on the Lexile scale; 320 on the Civics scale). Minimum scores, also known as cut-scores, are set through a 'standard-setting' process in which judgements are made item-by-item about the likely performance of a student who just satisfies the standard. For example, to set the

| | Literature Titles | Example Text |
|---|---|---|
| 990<br>980<br>960<br>960<br>920<br>920<br>**900** | *Jonathan Livingston Seagull*<br>*Fatherhood*<br>*The Adventures of Tom Sawyer*<br>*Pictionary Game Instructions*<br>*To Kill a Mockingbird*<br>*The Lion, the Witch & the Wardrobe* | I discussed the question in all its forms, politically and scientifically; and I give here an extract from a carefully studied article which I published in the number of the 30th of April. It ran as follows:– 'After examining one by one the different hypotheses, rejecting all other suggestions, it becomes necessary to admit the existence of a marine animal of enormous power. The great depths of the ocean are entirely unknown to us. Soundings cannot reach them …' |
| 890<br>880<br>870<br>830<br>810<br>810<br>**800** | *Stuart Little*<br>*The Red Pony*<br>*A Taste of Blackberries*<br>*Sounder*<br>*Mrs Frisby & Rats of NIMH*<br>*Johnny Appleseed* | It was higher than a big scythe blade and a very pale lavender above the dark blue water. It raked back and as the fish swam just below the surface the old man could see his huge bulk and the purple stripes that banded him. His dorsal fin was down and his huge pectorals were spread wide. On this circle the old man could see the fish's eye and the two grey sucking fish that swam around him. Sometimes they attached themselves to him. Sometimes they darted off. |
| 780<br>780<br>770<br>730<br>710<br>700<br>**700** | *Boy Scout Manual*<br>*Little House on the Prairie*<br>*The Cricket in Times Square*<br>*Harriet the Spy*<br>*Vanished*<br>*Where the Red Fern Grows* | Templeton, of course, was miserable over the loss of his beloved egg. But he couldn't resist boasting. 'It pays to save things,' he said in his surly voice. 'A rat never knows when something is going to come in handy. I never throw away anything.' 'Well,' said one of the lambs, 'this whole business is well and good for Charlotte, but what about the rest of us? The smell is unbearable. Who wants to live in a barn that is perfumed with rotten egg?' 'Don't worry, you'll get used to it,' said Templeton. |
| 690<br>670<br>650<br>640<br>620<br>610<br>**600** | *How to Eat Fried Worms*<br>*Chocolate Fever*<br>*On the Banks of Plum Creek*<br>*Hardy Boys Submarine Caper*<br>*Jack and Jill*<br>*Flossie and the Fox* | He did not know how the world is simplified for kings. To them, all men are subjects. 'Approach, so that I may see you better,' said the king, who felt consumingly proud of being at last a king over somebody. The little prince looked everywhere to find a place to sit down; but the whole planet was crammed and obstructed by the king's magnificent ermine robe. So he remained standing upright, and since he was tired, he yawned. 'It is contrary to etiquette to yawn in the presence of a king,' the monarch said. |
| 580<br>560<br>560<br>550<br>540<br>530<br>**500** | *The Whipping Boy*<br>*Something Over at Ballpark*<br>*Madeline's Rescue*<br>*The Boxcar Children*<br>*Sarah, Plain and Tall*<br>*A Boy in the Girls' Bathroom* | 'Aar.' Encyclopedia answered after a moment. He always waited a moment. He wanted to be helpful. But he was afraid that people might not like him if he answered their questions too quickly and sounded too smart. His father asked him more questions than anyone else. Mr Brown was the chief of police of Idaville. The town had four banks, three movie theatres, and a little league. It had the usual number of gasoline stations, churches, schools, stores, and comfortable houses on shady streets. |
| 490<br>480<br>450<br>450<br>440<br>430<br>**400** | *Commander Toad*<br>*Curious George*<br>*Alligator under my Bed*<br>*Sophie and Gussie*<br>*Something Queer Going On*<br>*Yonder* | The following Saturday morning my mother drove me to the highway to get the New York bus. It was my first time going alone and my mother was nervous. 'Listen, Margaret – don't sit next to any men. Either sit alone or pick out a nice lady. And try to sit up front. If the bus isn't air-conditioned open your window. And when you get there ask a lady to show you the way downstairs. Grandma will meet you at the information desk.' 'I know, I know.' We'd been over it three dozen times … |
| 380<br>370<br>350<br>320<br>310<br>300<br>**300** | *Tales of 4th Grade Nothing*<br>*Where is Cuddly Cat?*<br>*Little Rabbit*<br>*Adventures of Mr Toad*<br>*Ira Sleeps Over*<br>*Mog, the Forgetful Cat* | When Bear got home, he dumped all the money out of his piggy bank. Then he went downtown … and bought the moon a beautiful hat. That night he put the hat up in a tree where the moon could find it. Then he waited and watched while the moon slowly crept up through the branches and tried on the hat. 'Hurrah!' yelled Bear. 'It fits just right.' During the night, while bear slept, the hat fell out of the tree. In the morning Bear found the hat on his doorstep. 'So the moon got me a hat too!' exclaimed Bear. |
| 290<br>270<br>260<br>220<br>210<br>200<br>**200** | *Zack's Alligator*<br>*Bingo, Best Dog in the World*<br>*One Fish Two Red Fish Blue*<br>*Freight Train*<br>*Frog and Toad All Year*<br>*Ronald Morgan Goes to Bat* | In the great green room there was a telephone and a red balloon. And a picture of the cow jumping over the moon. And there were three little bears sitting on chairs. And two little kittens and a pair of mittens. And a little toyhouse and a young mouse. And a comb and a brush and a bowl full of mush. And a quiet old lady who was whispering 'hush'. Goodnight room. Goodnight moon. Goodnight cow jumping over the moon. Goodnight light and red balloon. Goodnight bears. Goodnight chairs. Goodnight kittens. |

pass score on a final-year Dentistry examination, experts in the field might judge the likelihood of a minimally competent dentist correctly answering each item on the examination.

When a performance standard is established, there is a special interest in knowing not only where students stand on an underlying continuum of achievement, but also where they stand in relation to a defined point (cut-score) on that continuum. In some contexts, such as final-year professional examinations, this question may be of primary interest in the interpretation of test results.

## Reporting Growth

A measurement variable also provides a frame of reference for monitoring and reporting growth over time. By measuring an individual's level of attainment on a variable on different occasions it is possible to track that individual's development over time, to plot his or her growth trajectory, and to evaluate improvement from one occasion to another. The interpretation of a student's current level of attainment by reference to that student's attainment on some earlier occasions is sometimes called 'ipsative' referencing.

In longitudinal studies of student achievement, the same individuals are tracked over a number of years. It is common in these studies to collect not only achievement measures, but also information about students' educational histories and experiences, their home backgrounds and relevant out-of-school activities. An attempt is then made to understand factors influencing learning over a number of years of school. Longitudinal studies depend on the possibility of making and comparing achievement measures on the same variable/s over extended age ranges.

It is also possible to measure and compare the achievements of a *group* of students on different occasions. When the progress of a group is followed, questions can be asked about average or typical rates of growth. The Third International Mathematics and Science Study (TIMSS), for example, measured the mathematics and science achievements of fourth-grade students in each participating country and, four years later, measured the achievements of eighth-grade students in those countries. In this way it was possible to measure average or typical growth in mathematics and science achievement over four years in each country.

## Comparing Attainments

Measures of student achievement also can be interpreted by comparing them with the achievements of other students. The process of comparing a student's measure with the measures of other students is known as 'norm referencing'.

A measure is interpreted 'normatively' whenever it is compared with the performances of others. The observations that a student has achieved the highest test score in her class, has performed in the top 10 per cent of students in the State, has a reading age of 6.2, and is achieving at the 85th percentile for her age group nationally are examples of norm-referenced interpretations of achievement.

If a student's achievement is to be interpreted by comparing it with the achievements of other students, then it is important to clarify the nature of the comparison group. Is the comparison group all 10-year-olds in the State/province? All 10-year-olds in the country? All fifth-grade students in the State/province? All fifth-graders in the country?

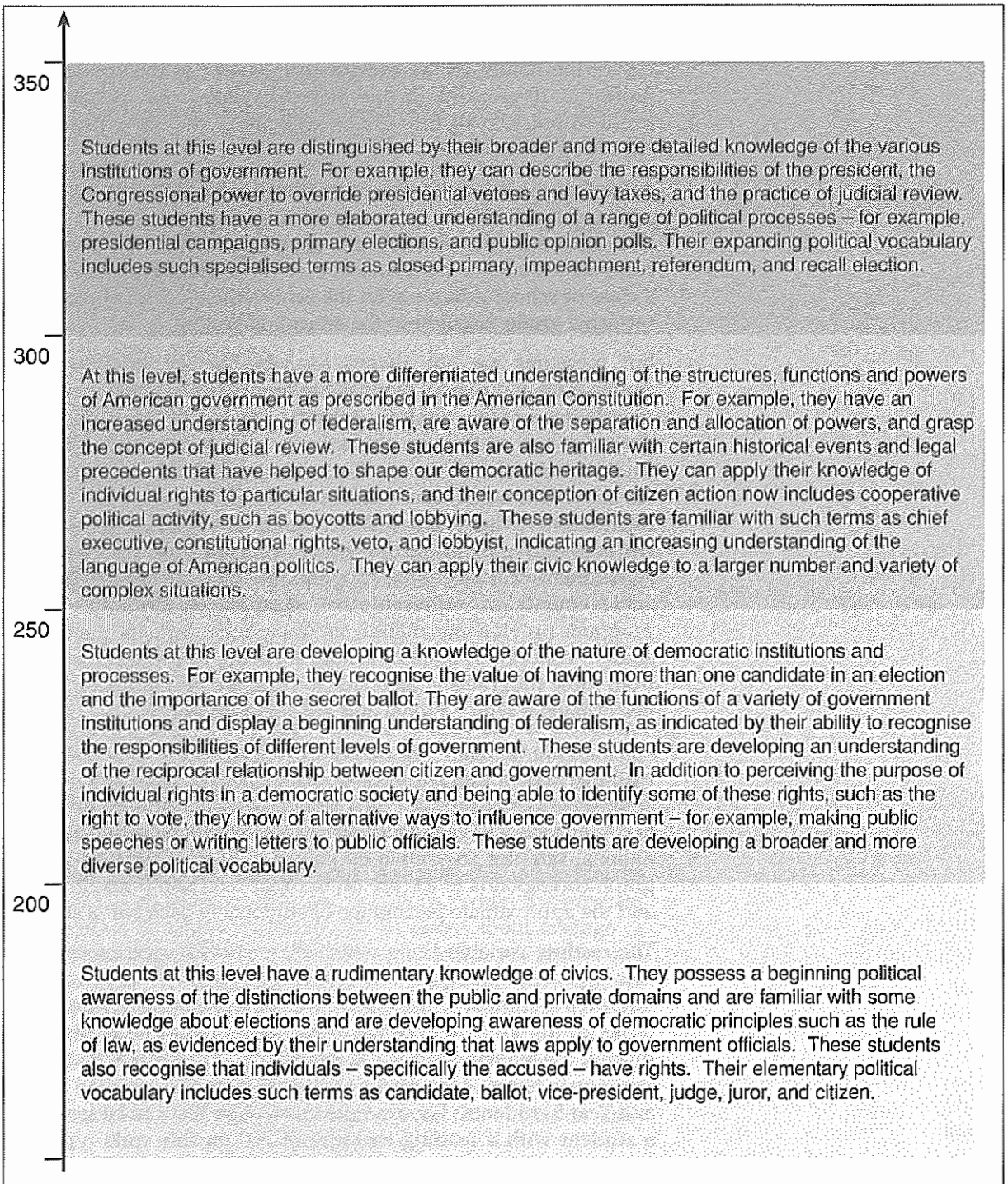
Some testing programs measure the achievements of all students in an education system at identified grade levels. These 'full-cohort' or 'population' testing programs make it possible to compare a student's performance – or the average performance of a class or school group – with the achievements of all students in the same grade throughout the education system.

But measures are not always available for all students in a comparison population, requiring inferences to be made about the population from a carefully selected sample of students. The drawing and testing of student samples is common practice in international achievement studies such as the Programme for International Student Assessment (PISA) and the Third International Mathematics and Science Study (TIMSS), and in national surveys of achievement such as the US National Assessment of Educational Progress (NAEP). By measuring the achievements of representative samples of students, these programs provide information about the achievements of national student cohorts. A State or school choosing to use test materials from these programs is then able to compare individual or group performances with national and international norms.

An example of a national survey of this kind was the Australian National School English Literacy Survey which measured the literacy achievements of carefully drawn national samples of Year 3 and Year 5 students. The measured reading achievements of these national samples are shown on page 42. Each of the bars in this graph corresponds to a score on the Year 3 or Year 5 reading test, and the approximate percentage of students in each bar is shown.

The reading variable along which these students were measured also appeared on page 30. By referring to page 30 it is possible to interpret a student's measured level of reading achievement in terms of the kinds of reading behaviours he or she is likely to display. By referring to page 42 it is possible to interpret that same measure in terms of the reading achievements of Australian Year 3 and Year 5 students. For example, from page 30 it can be seen that a student with a reading measure of 200 on this scale typically would be able to use combinations of pictures and text to demonstrate some understanding (eg, use a book title and cover illustration to identify key elements of a story; interpret a picture to predict what happens next in a story; use a title and illustration to predict a story setting; recognise how elements of an illustration support text in a story). From page 42 it can be seen that 89% of Year 3 students and 97% of Year 5 students were performing above this level.

Publishers of commercial tests usually provide test 'norms' allowing users to compare performances on a test with the performances of students of the same age or grade. Test norms show the percentage of students in a norming sample achieving each score on the test.




Described levels of attainment in Civics (US National Assessment of Educational Progress)

## Comparing Subgroups

When sufficiently large numbers of students are measured on the same variable, it is possible to compare and report the performances of student subgroups on that variable. The comparison of student subgroups is illustrated on page 43, where Year 3 and Year 5 students in the Australian National School English Literacy Survey have been grouped according to socio-economic status based on parents' occupations. Five

socio-economic categories were constructed at each Year level; the lowest (manual and unskilled labourers), middle, and highest (professional and managerial) are shown here. The box-and-whisker plots have been constructed to show the median, middle 60 per cent, and middle 80 per cent of students in each subgroup.

From the six box-and-whisker plots it can be seen that the difference between the reading achievements of the lowest and highest socio-economic groups is greater at Year 5 than at Year 3. It also can be seen that, on average, Year 3 students in the highest socio-economic group have higher reading levels than Year 5 students in the lowest group. It is common in statewide assessment programs and national and international surveys to identify student subgroups and to compare and report the achievements of these subgroups.

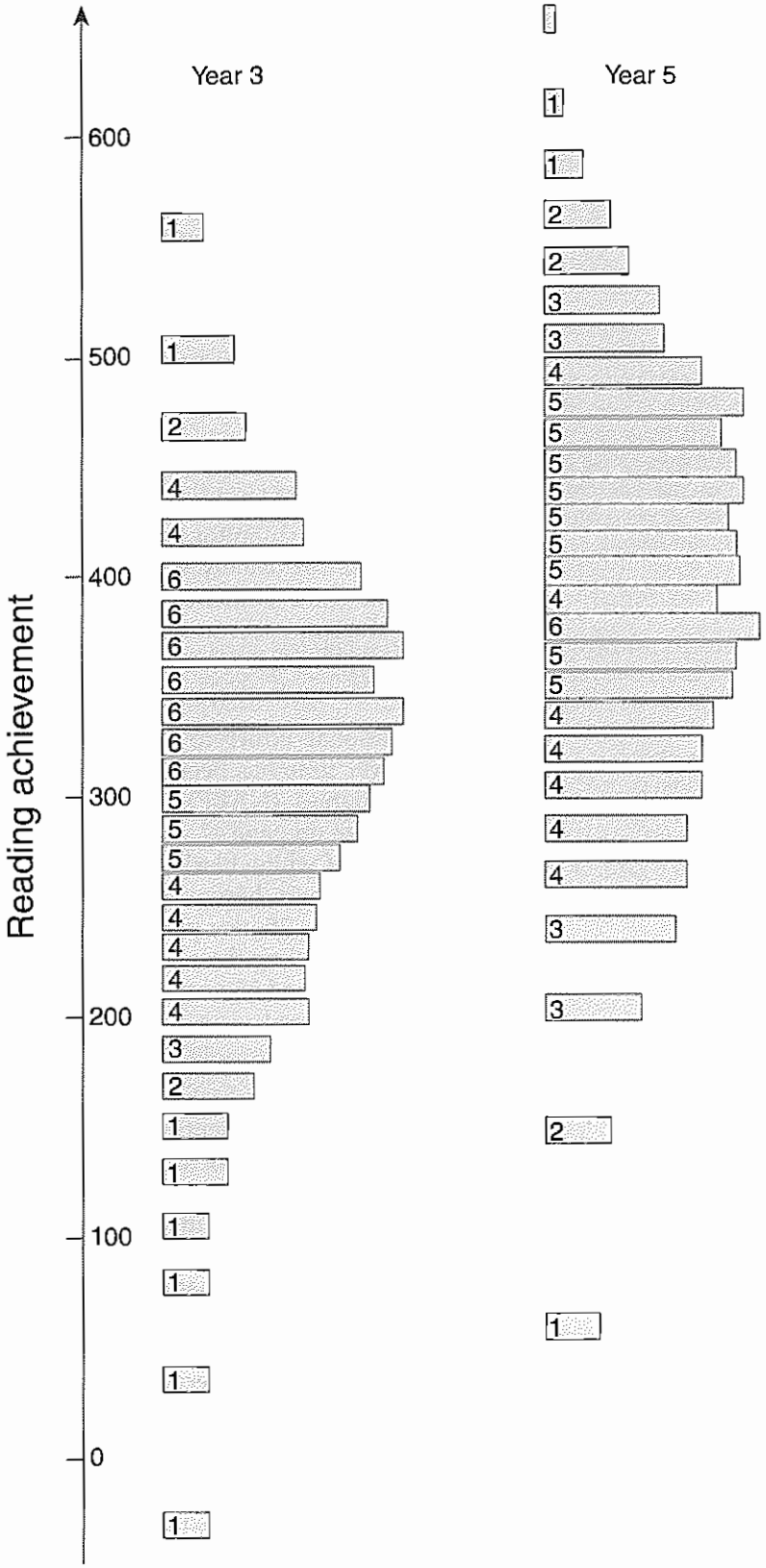
## Monitoring Trends Over Time

The construction and maintenance of measurement variables is essential to attempts to monitor trends in educational achievement over time. Because it usually is not desirable to administer the same test to the same students on different occasions, or even to administer the same test to different cohorts of students year after year, attempts to monitor trends over time depend on the calibration of different tests along a common variable.

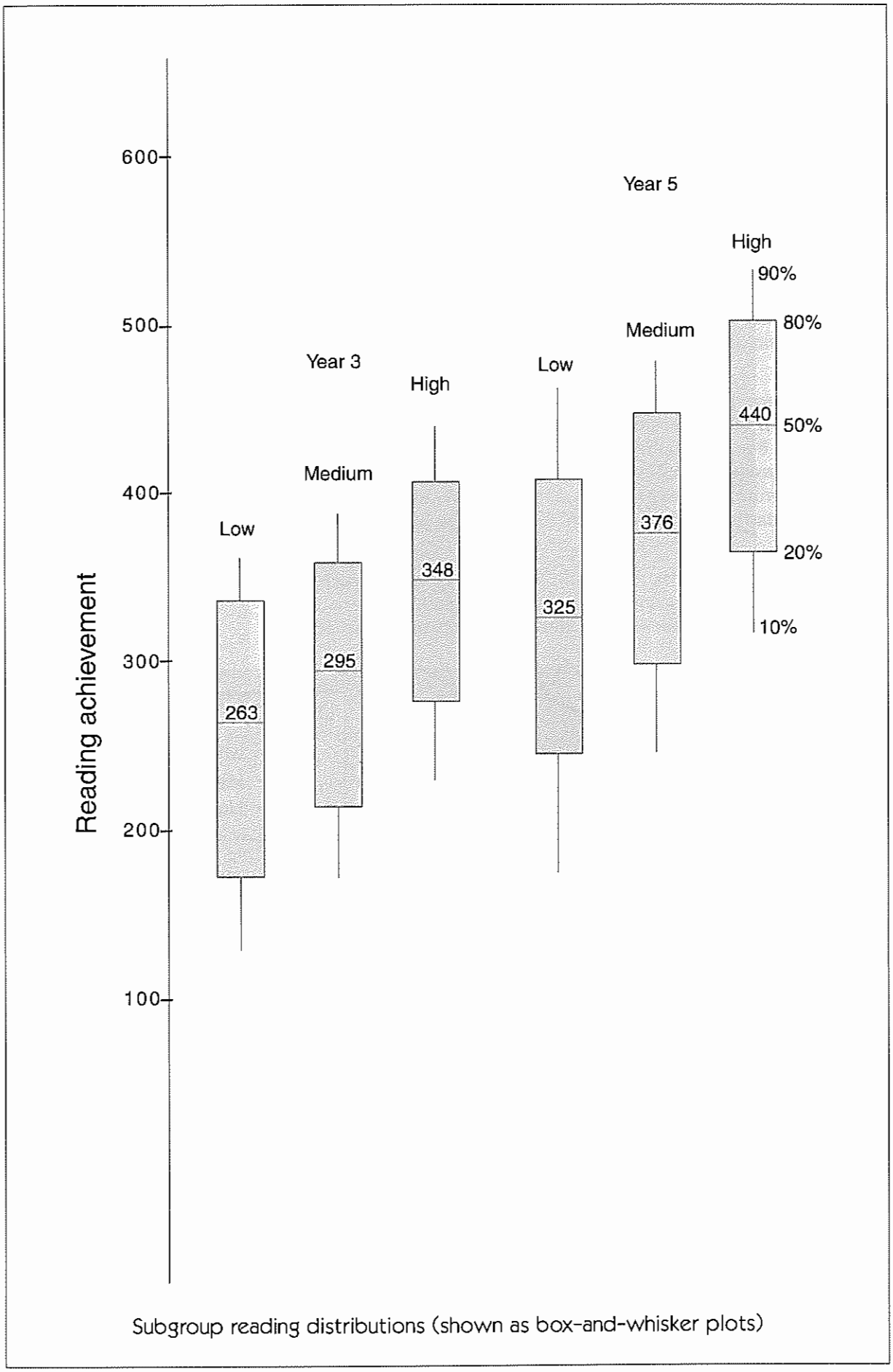
One major national effort to monitor trends in educational achievement over time has been the US National Assessment of Educational Progress (NAEP). Selected results from NAEP Reading are shown on page 44. Average reading achievements of national samples of white, black and Hispanic 9, 13 and 17-year-old students are shown here over the period 1971 to 1996. All reading tests over this 25-year period were calibrated on the same reading variable, enabling the reading levels of 9, 13 and 17-year-olds to be compared, and 25-year trends in reading achievement to be graphed and analysed.

From the graph on page 44 it can be seen that white students performed at significantly higher levels of reading achievement than black and Hispanic students throughout the 25-year period at all three ages. However, while there was no significant change in the reading levels of white students over this period, the average reading levels of black and Hispanic students increased significantly from 1971 to 1996. This was true at all three ages. Still closer inspection shows that, while there were improvements in the reading achievements of black and Hispanic students between 1971 and the late 1980s, there appears to have been no further improvement in the average achievement of these students during the 1990s.

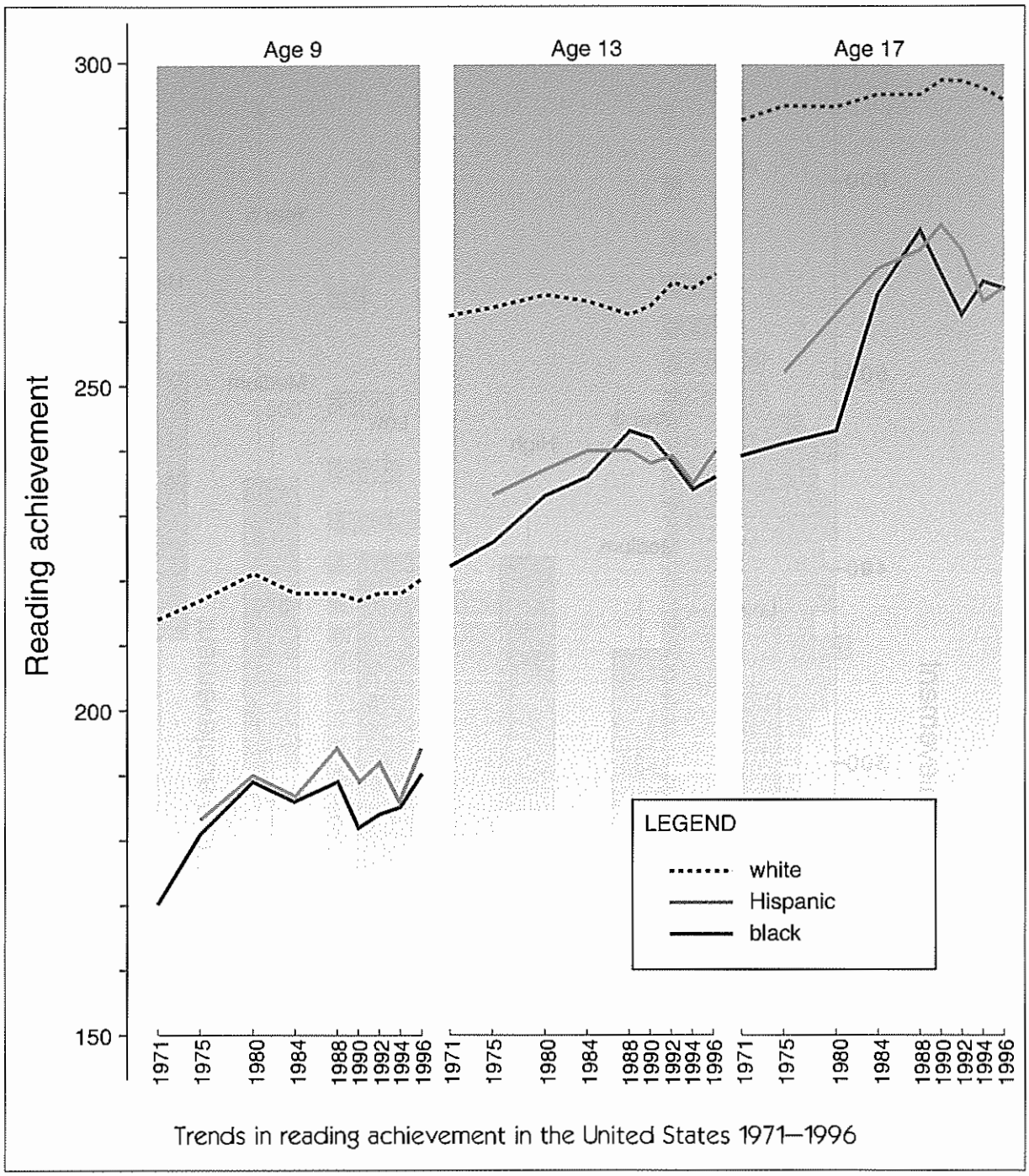
Many large-scale assessment programs, including a number of national and international surveys, provide education policy makers and administrators with information about trends in educational achievement. A prerequisite for the study of trends is a carefully constructed measurement variable along which growth and decline can be charted.




Distributions of Year 3 and Year 5 reading achievement




Subgroup reading distributions (shown as box-and-whisker plots)




## In Summary

Educational measurement is a process of estimating students' locations on some variable of interest. Once measures of an educational variable have been made, a variety of questions can be asked about any student's measure:

*What kinds of knowledge, skills, understandings, attitudes or values does the measure indicate?* In other words, where does the student stand on the variable of interest and what can be concluded about the student's current level of attainment? This question can be answered by referring to typical observations at varying locations along the variable (ie, the kinds of tasks the student is likely to be able to complete; the kinds of responses they are likely to give).

*Is the student performing at the level expected of students of his age or grade?* In other words, where does the student stand in relation to a pre-specified performance standard? This question can be answered by referring the student's measure to the expected or target level of attainment for the age/grade and deciding whether it is significantly above or below that level.

*What progress has the student made since the last assessment?* In other words, what growth has occurred? This question can be answered by measuring a student's achievement on a number of occasions and monitoring improvement over time.

*How does the student's attainment compare with the attainments of other students of the same age or grade?* In other words, how does it compare with age or grade norms? This question can be answered by referring the student's measure to the distribution of measures for the norm (reference) group.

Similar questions can be asked about entire groups of students. For example: What kinds of texts can be read and understood by the average six-year-old? What percentage of Year 5 students met the expected performance standard? What are typical rates of fine psychomotor skill development in 3-year-olds? How do national mathematics achievements compare with international benchmarks?

Answers to questions of this kind are essential to informed decision-making in education and depend on the availability of reliable measures on carefully constructed variables.

## What Is Measurement?

Measurement is the location of objects on variables by means of experience. We begin measurement with the idea of a variable. This idea can be visualised as a line pointing in a direction which indicates which way signifies 'more'. We give explicit meaning to a variable by specifying the kinds of questions (observations, test items) with which we hope to define it. We test the validity of these questions by exposing them to experience and discovering if there are useful circumstances under which they define a line. We make the meaning of the variable operational by estimating (calibrating) the relative positions of the valid questions on the line and labelling the line accordingly. If we can invent questions which retain their positions along a line over a useful range of applications, then we have a variable with an operational definition along which objective measurements can be attempted.

We make measurements on this variable by applying a suitable selection of questions to an object (person) we wish to measure, observing what happens and estimating from the pattern of reactions (responses) where among the ordered questions the object probably stands. We test the validity of this measurement by comparing the observed response pattern with its estimated expectation to see if the pattern can be accepted as a plausible outcome of the measuring system we have defined.

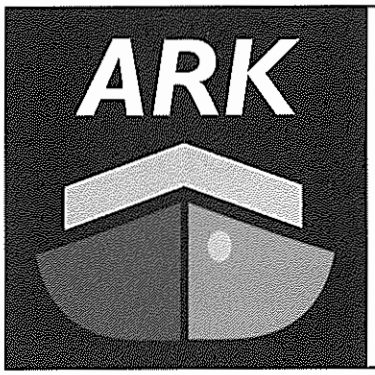
Benjamin D Wright
University of Chicago
30 March 1979

[1] Stenner, J (1996) *Measuring reading comprehension with the lexile framework.* Paper presented at the Fourth North American Conference on Adolescent/Adult Literacy. Washington.

[2] Anderson, L, Jenkins, LB, Leming, J, MacDonald, WB, Mullis, IVS, Turner, MJ & Wooster, JS (1990). *The Civics Report Card: Trends in Achievement from 1976 to 1988 at Ages 13 and 17.* Princeton: Education Testing Service.

# index

Please photocopy and fax or mail.




# ASSESSMENT RESOURCE KIT

## ORDER FORM

Australian Council for Educational Research
347 Camberwell Road (Private Bag 55)
Camberwell Victoria AUSTRALIA 3124
Telephone: (03) 9835 7447 Fax: (03) 9835 7499
International: Tel: 61 3 9835 7447 Fax: 61 3 9835 7499
Email: sales@acer.edu.au
Order online: www.acerpress.com.au

## Charge to

Name *or*
Organisation _____

Purchase Order no. *(if applicable)* _____

Address _____

_____ Postcode _____

## Deliver to

Name _____

Organisation _____

Street Address _____

_____ Postcode _____

Telephone ( ) _____ Fax ( ) _____

Order date _____ Date required _____ ACER account no. _____

| Title | Cat. No. | Price | Quantity | Total Price |
|---|---|---|---|---|
| Developmental Assessment | A100ARK | AUD$9.90 | | |
| Portfolios | A101ARK | AUD$9.90 | | |
| Performances | A102ARK | AUD$9.90 | | |
| Projects | A103ARK | AUD$9.90 | | |
| Progress Maps | A105ARK | AUD$9.90 | | |
| Products | A104ARK | AUD$9.90 | | |
| Assessment Methods | A106ARK | AUD$9.90 | | |
| Paper and Pen | A107ARK | AUD$9.90 | | |
| Videotape: Understanding Developmental Assessment | A700ARK | AUD$54.95 | | |
| Implementing Developmental Assessment: Workshop Manual & Videotape | A500ARK | AUD$38.50 | | |
| | | | SUBTOTAL | |

ADD FREIGHT:
*Within Australia:* 10% of invoice value, min. $5.50 max. $27.50 (inclusive of GST). Freight

NOTE: Prepayment required for non-account orders.
Prices subject to change without notice – you are welcome to check before you order. TOTAL $

☐ Please add my name and address to the mailing list for further information about ARK materials when they become available.

**CHARGE:** ☐ Account ☐ Cheque Enclosed (AUD$)

☐ Bankcard ☐ Mastercard ☐ American Express ☐ Visa ☐ Diners Club

ACER PRESS

☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐

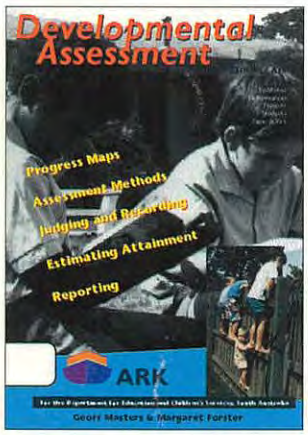
Name (*please print*) _____ Signature _____ Expiry date _____

ABN 004 398 145

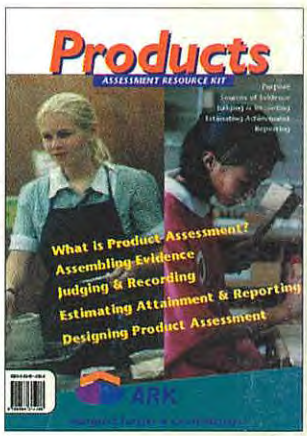
*Educational Measurement* is one in a series of magazines in the ACER Assessment Resource Kit (*ARK*).

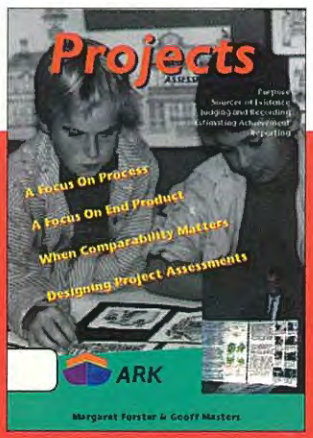
This video and magazine resource provides information about assessment issues and methods.

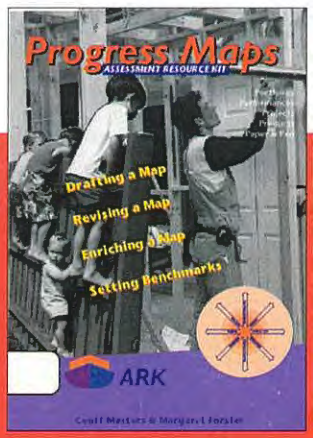
For further details about other magazines, videos and the workshop manual in this series contact the Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell, Victoria, Australia, 3124.
Phone: +61 3 9835 7447
Facsimile: +61 3 9835 7499

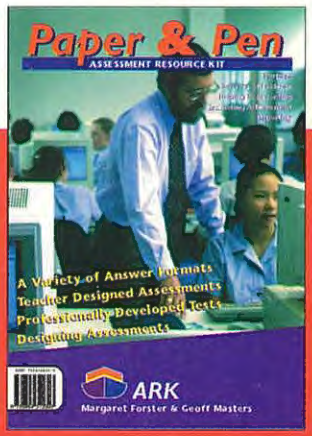





ACER