

EVALUATION IN EDUCATION

Dr. Sangeeta Agarwal



Evaluation in Education

Evaluation in Education

Dr. Sangeeta Agarwal (Associate Professor)
M.Ed., Ph.D. (Education), M.Com., M.A. (English),
Faculty of Education, Tania University, Sri Ganganagar



Evaluation in Education

Dr. Sangeeta Agarwal

© RESERVED

This book contains information obtained from highly regarded resources. Copyright for individual articles remains with the authors as indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereinafter invented, including photocopying, microfilming and recording, or any information storage or retrieval system, without permission from the publishers.



4378/4-B, Murarilal Street, Ansari Road, Daryaganj, New Delhi-110002.
Ph. No: +91-11-23281685, 41043100, Fax: +91-11-23270680
E-mail: academicuniversitypress@gmail.com

Year of Publication 2024-25

ISBN : 978-93-6284-007-3

Printed and bound by: Global Printing Services, Delhi
10 9 8 7 6 5 4 3 2 1

Contents

<i>Preface</i>	<i>vii</i>
Chapter 1 Introduction	1
Chapter 2 Educational Measurement	26
Chapter 3 Measurement, Evaluation and Research	50
Chapter 4 Test Assessment	80
Chapter 5 Special Measurement Techniques	102
Chapter 6 Evaluation in a Democratic School	136

Preface

Evaluation in education is a critical process that involves assessing the effectiveness, efficiency, and quality of educational programmes, policies, and practices. It encompasses a systematic approach to gathering and analyzing data to determine the extent to which educational objectives are being met and to identify areas for improvement. Evaluation serves multiple purposes, including informing decision-making, providing accountability, and promoting continuous improvement in educational settings. Through evaluation, educators and policymakers can assess the impact of educational interventions, identify best practices, and allocate resources effectively to support student learning and success.

The process of evaluation in education involves several key steps, beginning with the formulation of clear and measurable goals and objectives. These goals provide a framework for assessing the outcomes and effectiveness of educational programmes. Evaluation methods and instruments are then selected based on the goals and objectives, which may include qualitative and quantitative approaches such as surveys, interviews, observations, and standardized assessments. Data is collected through these methods and analyzed to determine the strengths and weaknesses of the programme or practice being evaluated.

Furthermore, evaluation in education requires careful consideration of ethical and cultural factors to ensure that assessments are fair, unbiased, and culturally responsive. It is essential to involve stakeholders, including students, parents, educators, and community members, in the evaluation process to gather diverse perspectives and insights. Collaboration and communication among stakeholders are critical for building consensus, addressing concerns, and fostering a shared understanding of evaluation findings.

Moreover, the results of evaluation in education are used to inform decision-making, policy development, and programme improvement. By analyzing evaluation data, educators and policymakers can identify areas of success and areas in need of improvement, leading to targeted interventions and adjustments to educational practices. Evaluation also plays a crucial role in accountability, as it provides evidence of the impact and effectiveness of educational investments and initiatives.

This book evaluation in education is a dynamic and multifaceted process that is essential for promoting accountability, driving improvement, and enhancing the quality of educational programmes and practices. By systematically assessing the outcomes and effectiveness of educational interventions, evaluation helps to ensure that resources are allocated efficiently and that all students have access to high-quality education that meets their needs and supports their success.

–Author

1

Introduction

Measurement, assessment, and evaluation mean very different things, and yet most of my students were unable to adequately explain the differences. Measurement refers to the process by which the attributes or dimensions of some physical object are determined. One exception seems to be in the use of the word measure in determining the IQ of a person. The phrase, “this test measures IQ” is commonly used. Measuring such things as attitudes or preferences also applies. However, when we measure, we generally use some standard instrument to determine how big, tall, heavy, voluminous, hot, cold, fast, or straight something actually is. Standard instruments refer to physical devices such as rulers, scales, thermometers, pressure gauges, *etc.*

We measure to obtain information about what is. Such information may or may not be useful, depending on the accuracy of the instruments we use, and our skill at using them. Assessment is a process by which information is obtained relative to some known objective or goal. Assessment is a broad term that includes testing. A test is a special form of assessment. Tests are assessments made under contrived circumstances especially so that they may be administered. In other words, all tests are assessments, but not all assessments are tests. Evaluation is perhaps the most complex and least understood of the terms. Inherent in the idea of evaluation is “value.” When we evaluate, what we are doing is engaging in some process that is designed to provide information that will help us make a judgement about a given situation. Generally, any evaluation process requires information about the situation in question. A situation is an umbrella term that takes into account such ideas as objectives, goals, standards, procedures, and so on. When we evaluate, we are saying that the process will

yield information regarding the worthiness, appropriateness, goodness, validity, legality, *etc.*, of something for which a reliable measurement or assessment has been made.

EDUCATIONAL EVALUATION

Educational evaluation is the evaluation process of characterising and appraising some aspect/s of an educational process.

There are two common purposes in educational evaluation which are, at times, in conflict with one another. Educational institutions usually require evaluation data to demonstrate effectiveness to funders and other stakeholders, and to provide a measure of performance for marketing purposes. Educational evaluation is also a professional activity that individual educators need to undertake if they intend to continuously review and enhance the learning they are endeavouring to facilitate.

Standards for Educational Evaluation

The Joint Committee on Standards for Educational Evaluation published three sets of standards for educational evaluations. *The Personnel Evaluation Standards* was published in 1988, *The Programme Evaluation Standards* (2nd edition) was published in 1994, and *The Student Evaluations Standards* was published in 2003.

Each publication presents and elaborates a set of standards for use in a variety of educational settings. The standards provide guidelines for designing, implementing, assessing and improving the identified form of evaluation. Each of the standards has been placed in one of four fundamental categories to promote evaluations that are proper, useful, feasible, and accurate.

THE PERSONNEL EVALUATION STANDARDS

- The propriety standards require that evaluations be conducted legally, ethically, and with due regard for the welfare of evaluatees and clients involved in.
- The utility standards are intended to guide evaluations so that they will be informative, timely, and influential.
- The feasibility standards call for evaluation systems that are as easy to implement as possible, efficient in their use of time and resources, adequately funded, and viable from a number of other standpoints.
- The accuracy standards require that the obtained information be technically accurate and that conclusions be linked logically to the data.

The Programme Evaluation Standards

- The utility standards are intended to ensure that an evaluation will serve the information needs of intended users.

- The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.
- The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.
- The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the programme being evaluated.

3.4.4 The Student Evaluation Standards

- The Propriety standards help ensure that student evaluations are conducted lawfully, ethically, and with regard to the rights of students and other persons affected by student evaluation.
- The Utility standards promote the design and implementation of informative, timely, and useful student evaluations.
- The Feasibility standards help ensure that student evaluations are practical; viable; cost-effective; and culturally, socially, and politically appropriate.
- The Accuracy standards help ensure that student evaluations will provide sound, accurate, and credible information about student learning and performance.

THE IMPORTANCE OF EDUCATIONAL MEASUREMENT AND EVALUATION

As much as we might want to embrace the “everyone’s a winner” mentality, evaluation is a crucial component of education. Without evaluation and measurement, it is impossible to know a student’s needs and preferences. Evaluation is also used by colleges to determine which students can be admitted. While the specific purposes of measurement and evaluation can vary, there is one underlying theme: measurement and evaluation are required to determine whether students are learning.

Student Needs

The most basic purpose of educational evaluation is to determine what a student’s needs are. With proper testing and evaluation in the early grades, learning disabilities and handicaps can be identified and dealt with. Without testing, problems can go unrecognised for years. While educational testing cannot in itself be the basis for a diagnosis, it can point students in a direction that may ultimately lead to psychologist, who can diagnose conditions.

Student Aptitudes

In the 21st century, there is much emphasis on specialisation in education. In today’s complex, knowledge-based economy, students must have specialised skills before they can have a successful career. The streaming of students into educational programmes begins with standardised testing, which identifies

student aptitudes and abilities. While standardised tests are somewhat controversial due to their potential for misuse, there is no denying that they can be effective in identifying intellectual gifts and helping students know the areas in which their talents can be useful.

Student Progress

Education is effective when students improve over time. Without measurement and evaluation, it is impossible to know whether students are making any progress. Tests and assignments can tell teachers which students know the material, which students are trying to learn and which students are not trying at all.

While evaluations are not perfect in determining student achievement (some students underperform in spite of effort because of learning disabilities), the progress in a student's grades over time can say a lot about where that student is and where he needs to be.

College Admissions

Eventually, students get to their last year of high school. At this point, a student needs to score the best grades possible in order to get into a good college. While it is possible to debate the merits of college admissions processes based primarily on grades, there is no question that such a process is in place at many colleges. Therefore, a rigorous system of testing and evaluation, in which teachers provide students with many smaller assignments in order to address issues before the big assignments, will help a school's college placement rate in the long run.

THE FUNCTIONS OF MEASUREMENT AND EVALUATION IN IMPROVING INSTRUCTION

Teaching, learning and evaluation are three interdependent aspects of the educative process. Therefore, evaluation is an indispensable part of the teaching-learning process.

It involves measurement and assigning qualitative meaning through value judgements. It is a means of determining the effectiveness of teaching methodologies, instructional materials and other elements affecting the teaching-learning situation.

The importance of evaluation cannot be overemphasized. It is important as an instrument of the school system, to the teacher, the learner, the parent and the administrator for the improvement of instruction.

Evaluation involves the determination of the goals and objectives towards which educational efforts are directed and the determination of measurement techniques to be utilised in the assessment of desired goals and objectives. It includes the assessment of all elements of the teaching-learning situation that contribute to the effective learning with the end in view of improvement.

Through assessment and evaluation, pupils' achievement interest, success, difficulty and instruction can be assessed properly. The result of evaluation can be used as a benchmark for instructional enhancement.

The following are the functions of measurement and evaluation in improving instruction.

1. Evaluation results enable the teacher to accumulate the experiences and to follow-up diagnosed results. The weaknesses of the pupils in the class can be identified and remedied, thus pupils' performance is enhanced.
2. Measurement and evaluation measure pupils' achievement and motivate pupils' learning. Pupils have the right to know the progress they are making whether they have attained the objectives of the subject matter or not, thus results must be made known to them. It can also encourage pupils to study more. They will be motivated to participate actively in class and exert all efforts just to make certain that they pass. They will know the quality and amount of work they have to strive for.
3. Measurement and evaluation predict pupils' success and diagnoses pupils' difficulty. The success and failure of a pupil in the class can be predicted through it. The area where pupils excel must be enhanced or strengthened and where pupils fail should be remedied. The difficulties of the pupils should be given the priority for remediation. Knowing the successes and difficulties of the pupils, the teacher will be able to focus on the spots that need enhancement or remediation.

It is hope that evaluation results serve as the basis for the teacher to use appropriate teaching strategies and techniques that will improve instruction and provide the necessary learning for a pupil to acquire the knowledge and skills he needs.

THE DIFFERENCE BETWEEN EVALUATION, MEASUREMENT, TEST AND ASSESSMENT

Many people confuse about the notion of evaluation, some of them didn't understand the different between measurement, test, and assessment. Did You know it? Here I will give a little explanation about that.

Evaluation is an identification activity to see if a programme that has been planned achieved or not, valuable or not. It's also to see the level of implementation efficiency. Evaluation is relating to the decision value. Stufflebeam said that: *educational evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives.* From Stufflebeam views, we can see that the essence of evaluation is to provide information for decision making purposes.

In education, we can evaluate the new curriculum, an education policy, specific learning resources, teacher or work ethic. The test is the assessment ways to designed and implemented to students at a particular time and place

and under conditions that meet certain requirements are clear. In particular, in the context of learning in the classroom, the assessment is to determine the progress and outcomes of students, diagnosing learning difficulties, provide feedback/improvement of teaching and learning process, and determining the increase in class. The assessment can be obtained accurate information about the organisation of teaching and learning success of students, teachers, and the learning process itself. Based on that information, teacher can make decisions about learning, learner's difficulties and the effort necessary guidance and presence of curriculum itself.

OBJECTIVES ASSESSMENT

Assessment has a very important purpose in learning, such as for grading, selection, knowing the level of mastery of competencies, counselling, diagnosis, and prediction.

1. For grading, the *assessment* is intended to define or distinguish the position of student work compared with other learners. This assessment will indicate the position of learners in the sequence compared with other children. Therefore, the function of assessment for grading is likely to compare the child with the other children so that more refer to the assessment of the reference norm (norm-referenced assessment).
2. As a means of selection, assessment is intended to separate between learners who fall into certain categories and who do not. Learners are allowed to enter a particular school or who are not allowed. In this case, the function of assessment to determine a person can enter or not in a particular school.
3. To illustrate the extent to which a student has mastered the competencies.
4. As guidance, the assessment aims to evaluate the learning outcomes of learners in order to help learners understand themselves, make decisions about next steps, both for the selection of courses, personality development as well as for the majors.
5. As a means of diagnosis, the assessment aims to show learning difficulties experienced by learners and the possibility of achievement that can be developed. This will help teachers determine whether a person needs to remedies or enrichment.
6. As a predictive tool, the assessment aims to obtain information that can predict how the performance of students in the next education level or in a suitable job. Examples of this assessment are the scholastic aptitude tests or tests of academic potential.

From the sixth assessment objective, the goal is to see the level mastery of competencies, mentoring, and also a major role in the assessment, diagnostics.

In accordance with these goals, the assessment requires teachers to directly or indirectly capable of conducting an assessment in the overall learning process. To assess the extent students have mastered a variety of competencies, of course various types of assessment needs to be given in accordance with the

competencies to be assessed, such as performance, assignment (project), the work (product), a collection of student work (portfolio), and a written assessment (paper and pencil test). Thus, the purpose of assessment is to provide comprehensive input information about learners' learning outcome, whether viewed as the learning activities take place and when viewed from the end result, using various ways in accordance with the competency assessment that is expected to reach learners.

PSYCHOMETRICS

Psychometrics is the field of study concerned with the theory and technique of psychological measurement, which includes the measurement of knowledge, abilities, attitudes, personality traits, and educational measurement. The field is primarily concerned with the construction and validation of measurement instruments such as questionnaires, tests, and personality assessments.

It involves two major research tasks, namely: (i) the construction of instruments and procedures for measurement; and (ii) the development and refinement of theoretical approaches to measurement. Those who practice psychometrics are known as psychometricians. All psychometricians possess a specific psychometric qualification, and while many are clinical psychologists, others work as human resources or learning and development professionals.

19th Century Foundation

Psychological testing has come from two streams of thought: one, from Darwin, Galton, and Cattell on the measurement of individual differences, and the second, from Herbart, Weber, Fechner, and Wundt and their psychophysical measurements of a similar construct. The second set of individuals and their research is what has led to the development of experimental psychology, and standardised testing.

Victorian Stream

Charles Darwin was the inspiration behind Sir Francis Galton who led to the creation of psychometrics. In 1859, Charles Darwin published his book "The Origin of Species", which pertained to individual differences in animals. This book discussed how individual members in a species differ and how they possess characteristics that are more adaptive and successful or less adaptive and less successful. Those who are adaptive and successful are the ones that survive and give way to the next generation, who would be just as or more adaptive and successful. This idea, studied previously in animals, led to Galton's interest and study of human beings and how they differ one from another, and more importantly, how to measure those differences.

Galton wrote a book entitled "Hereditary Genius" about different characteristics that people possess and how those characteristics make them more "fit" than others. Today these differences, such as sensory and motor functioning (reaction time, visual acuity, and physical strength) are important domains of scientific psychology. Much of the early theoretical and applied work in psychometrics

was undertaken in an attempt to measure intelligence. Francis Galton, often referred to as “the father of psychometrics,” devised and included mental tests among his anthropometric measures. James McKeen Cattell, who is considered a pioneer of psychometrics went on to extend Galton’s work. Cattell also coined the term *mental test*, and is responsible for the research and knowledge which ultimately led to the development of modern tests.

German Stream

The origin of psychometrics also has connections to the related field of psychophysics. Around the same time that Darwin, Galton, and Cattell were making their discoveries, J.E. Herbart was also interested in “unlocking the mysteries of human consciousness” through the scientific method. Herbart was responsible for creating mathematical models of the mind, which were influential in educational practices in years to come.

Following Herbart, E.H. Weber built upon Herbart’s work and tried to prove the existence of a psychological threshold saying that a minimum stimulus was necessary to activate a sensory system. After Weber, G.T. Fechner expanded upon the knowledge he gleaned from Herbart and Weber, to devise the law that the strength of a sensation grows as the logarithm of the stimulus intensity. A follower of Weber and Fechner, Wilhelm Wundt is credited with founding the science of psychology. It is Wundt’s influence that paved the way for others to develop psychological testing.

20th Century

The psychometrician L. L. Thurstone, founder and first president of the Psychometric Society in 1936, developed and applied a theoretical approach to measurement referred to as the law of comparative judgement, an approach that has close connections to the psychophysical theory of Ernst Heinrich Weber and Gustav Fechner. In addition, Spearman and Thurstone both made important contributions to the theory and application of factor analysis, a statistical method developed and used extensively in psychometrics. In the late 1950s, Leopold Szondi made an historical and epistemological assessment of the impact of statistical thinking onto psychology during previous few decades: “in the last decades, the specifically psychological thinking has been almost completely suppressed and removed, and replaced by a statistical thinking. Precisely here we see the cancer of testology and testomania of today.”

More recently, psychometric theory has been applied in the measurement of personality, attitudes, and beliefs, and academic achievement. Measurement of these unobservable phenomena is difficult, and much of the research and accumulated science in this discipline has been developed in an attempt to properly define and quantify such phenomena. Critics, including practitioners in the physical sciences and social activists, have argued that such definition and quantification is impossibly difficult, and that such measurements are often misused, such as with psychometric personality tests used in employment procedures:

“For example, an employer wanting someone for a role requiring consistent attention to repetitive detail will probably not want to give that job to someone who is very creative and gets bored easily.”

Figures who made significant contributions to psychometrics include Karl Pearson, Henry F. Kaiser, Carl Brigham, L. L. Thurstone, Georg Rasch, Eugene Galanter, Johnson O'Connor, Frederic M. Lord, Ledyard R Tucker, Arthur Jensen, and David Andrich. Psychometric, psychometrician and psychometrist appreciation week is the first week in November.

THE RASCH MODEL FOR MEASUREMENT

In the Rasch model, the probability of a specified response (*e.g.*, right/wrong answer) is modelled as a function of person and item parameters. Specifically, in the original Rasch model, the probability of a correct response is modelled as a logistic function of the difference between the person and item parameter. The mathematical form of the model is provided later in this article. In most contexts, the parameters of the model characterise the proficiency of the respondents and the difficulty of the items as locations on a continuous latent variable. For example, in educational tests, item parameters represent the difficulty of items while person parameters represent the ability or attainment level of people who are assessed. The higher a person's ability relative to the difficulty of an item, the higher the probability of a correct response on that item. When a person's location on the latent trait is equal to the difficulty of the item, there is by definition a 0.5 probability of a correct response in the Rasch model. A Rasch model is a *model* in one sense in that it represents the structure which data should exhibit in order to obtain measurements from the data; *i.e.*, it provides a criterion for successful measurement. Beyond data, Rasch's equations model relationships we expect to obtain in the real world. For instance, education is intended to prepare children for the entire range of challenges they will face in life, and not just those that appear in textbooks or on tests. By requiring measures to remain the same (invariant) across different tests measuring the same thing, Rasch models make it possible to test the hypothesis that the particular challenges posed in a curriculum and on a test coherently represent the infinite population of all possible challenges in that domain. A Rasch model is therefore a model in the sense of an *ideal* or standard that provides a heuristic fiction serving as a useful organising principle even when it is never actually observed in practice.

The perspective or paradigm underpinning the Rasch model is distinct from the perspective underpinning statistical modelling. Models are most often used with the intention of describing a set of data. Parameters are modified and accepted or rejected based on how well they fit the data. In contrast, when the Rasch model is employed, the objective is to obtain data which fit the model (Andrich, 2004; Wright, 1984, 1999). The rationale for this perspective is that the Rasch model embodies requirements which must be met in order to obtain measurement, in the sense that measurement is generally understood in the

physical sciences. A useful analogy for understanding this rationale is to consider objects measured on a weighing scale. Suppose the weight of an object A is measured as being substantially greater than the weight of an object B on one occasion, then immediately afterwards the weight of object B is measured as being substantially greater than the weight of object A. A property we require of measurements is that the resulting comparison between objects should be the same, or invariant, irrespective of other factors. This key requirement is embodied within the formal structure of the Rasch model. Consequently, the Rasch model is not altered to suit data. Instead, the method of assessment should be changed so that this requirement is met, in the same way that a weighing scale should be rectified if it gives different comparisons between objects upon separate measurements of the objects. Data analysed using the model are usually responses to conventional items on tests, such as educational tests with right/wrong answers. However, the model is a general one, and can be applied wherever discrete data are obtained with the intention of measuring a quantitative attribute or trait.

Scaling

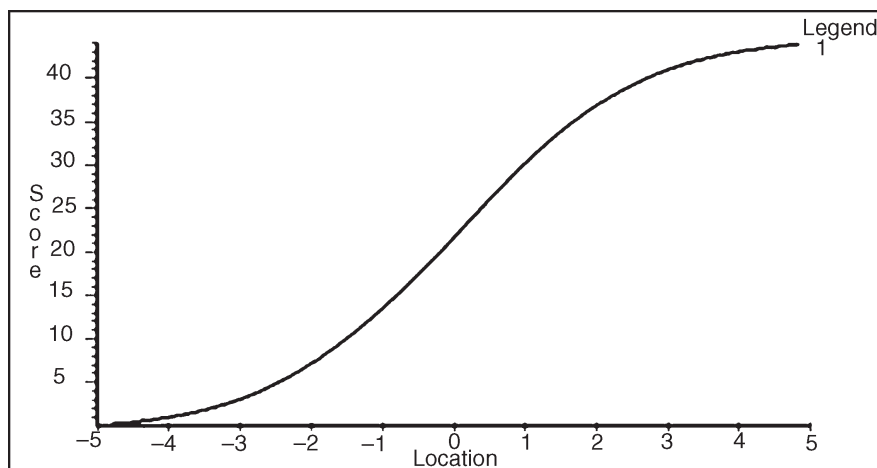


Fig. Test Characteristic Curve Showing the Relationship Between Total Score on a Test and Person Location Estimate.

When all test-takers have an opportunity to attempt all items on a single test, each total score on the test maps to a unique estimate of ability and the greater the total, the greater the ability estimate. Total scores do not have a linear relationship with ability estimates. Rather, the relationship is non-linear. The total score is shown on the vertical axis, while the corresponding person location estimate is shown on the horizontal axis. For the particular test on which the test characteristic curve (TCC) shown in Figure is based, the relationship is approximately linear throughout the range of total scores from about 10 to 33. The shape of the TCC is generally somewhat sigmoid as in this example. However, the precise relationship between total scores and person location

estimates depends on the distribution of items on the test. The TCC is steeper in ranges on the continuum in which there are a number of items, such as in the range on either side of 0 in Figures. In applying the Rasch model, item locations are often scaled first, based on methods such as those described below. This part of the process of scaling is often referred to as item *calibration*. In educational tests, the smaller the proportion of correct responses, the higher the difficulty of an item and hence the higher the item's scale location. Once item locations are scaled, the person locations are measured on the scale. As a result, person and item locations are estimated on a single scale.

Interpreting Scale Locations

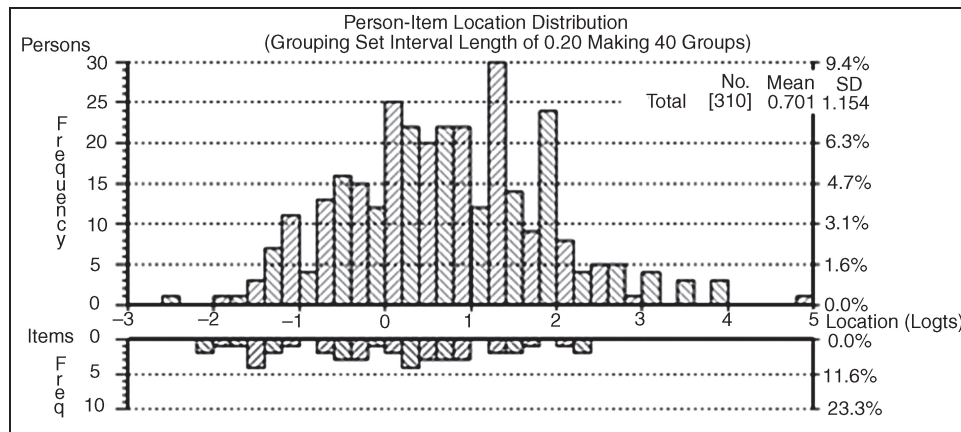


Fig. Graph Showing Histograms of Person Distribution (top) and Item Distribution (Bottom) on a Scale

For dichotomous data such as right/wrong answers, by definition, the location of an item on a scale corresponds with the person location at which there is a 0.5 probability of a correct response to the question. In general, the probability of a person responding correctly to a question with difficulty lower than that person's location is greater than 0.5, while the probability of responding correctly to a question with difficulty greater than the person's location is less than 0.5. The Item Characteristic Curve (ICC) or Item Response Function (IRF) shows the probability of a correct response as a function of the ability of persons. A single ICC is shown and explained in more detail in relation to Figure below in this article. The leftmost ICCs in Figure below are the easiest items, the rightmost items in the same figure are the most difficult items. When responses of a person are listed according to item difficulty, from lowest to highest, the most likely pattern is a Guttman pattern or vector; *i.e.*, $\{1,1,...,1,0,0,0,...,0\}$. However, while this pattern is the most probable given the structure of the Rasch model, the model requires only probabilistic Guttman response patterns; that is, patterns which tend towards the Guttman pattern. It is unusual for responses to conform strictly to the pattern because there are many possible patterns. It is unnecessary for responses to conform strictly to the pattern in order for data to fit the Rasch model.

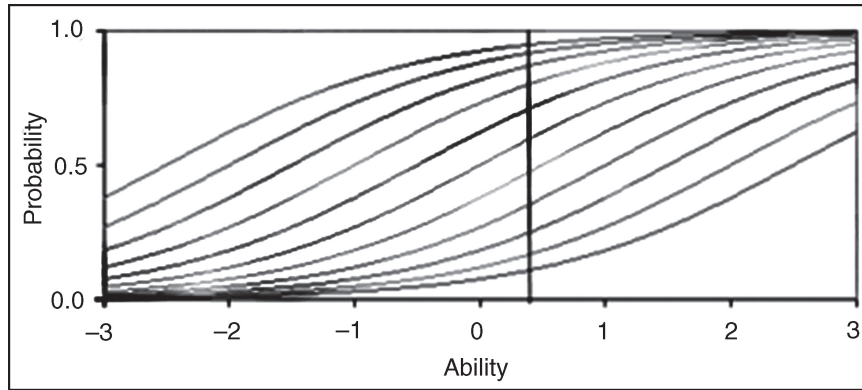


Fig. ICCs for a Number of Items. ICCs are Coloured to Highlight the Change in the Probability of a Successful Response for a Person with Ability Location at the Vertical line.

The person is likely to respond correctly to the easiest items (with locations to the left and higher curves) and unlikely to respond correctly to difficult items (locations to the right and lower curves).

Each ability estimate has an associated standard error of measurement, which quantifies the degree of uncertainty associated with the ability estimate. Item estimates also have standard errors. Generally, the standard errors of item estimates are considerably smaller than the standard errors of person estimates because there are usually more response data for an item than for a person. That is, the number of people attempting a given item is usually greater than the number of items attempted by a given person. Standard errors of person estimates are smaller where the slope of the ICC is steeper, which is generally through the middle range of scores on a test. Thus, there is greater precision in this range since the steeper the slope, the greater the distinction between any two points on the line.

Statistical and graphical tests are used to evaluate the correspondence of data with the model. Certain tests are global, while others focus on specific items or people. Certain tests of fit provide information about which items can be used to increase the reliability of a test by omitting or correcting problems with poor items. In Rasch Measurement the person separation index is used instead of reliability indices. However, the person separation index is analogous to a reliability index. The separation index is a summary of the genuine separation as a ratio to separation including measurement error. As mentioned earlier, the level of measurement error is not uniform across the range of a test, but is generally larger for more extreme scores (low and high).

Features of the Rasch Model

The class of models is named after Georg Rasch, a Danish mathematician and statistician who advanced the epistemological case for the models based on their congruence with a core requirement of measurement in physics; namely the requirement of *invariant comparison*. This is the defining feature of the class of models, as is elaborated upon in the following section. The Rasch model for dichotomous data has

a close conceptual relationship to the law of comparative judgement (LCJ), a model formulated and used extensively by L. L. Thurstone, and therefore also to the Thurstone scale. Prior to introducing the measurement model he is best known for, Rasch had applied the Poisson distribution to reading data as a measurement model, hypothesising that in the relevant empirical context, the number of errors made by a given individual was governed by the ratio of the text difficulty to the person's reading ability. Rasch referred to this model as the *multiplicative Poisson model*. Rasch's model for dichotomous data – *i.e.*, where responses are classifiable into two categories — is his most widely known and used model, and is the main focus here. This model has the form of a simple logistic function.

The brief outline above highlights certain distinctive and interrelated features of Rasch's perspective on social measurement, which are as follows:

1. He was concerned principally with the measurement of *individuals*, rather than with distributions among populations.
2. He was concerned with establishing a basis for meeting a priori *requirements* for measurement deduced from physics and, consequently, did not invoke any *assumptions* about the distribution of levels of a trait in a population.
3. Rasch's approach explicitly recognises that it is a scientific hypothesis that a given trait is both quantitative and measurable, as operationalised in a particular experimental context.

Thus, congruent with the perspective articulated by Thomas Kuhn in his 1961 paper *The function of measurement in modern physical science*, measurement was regarded both as being founded in theory, and as being instrumental to detecting quantitative anomalies incongruent with hypotheses related to a broader theoretical framework. This perspective is in contrast to that generally prevailing in the social sciences, in which data such as test scores are directly treated as measurements without requiring a theoretical foundation for measurement. Although this contrast exists, Rasch's perspective is actually complementary to the use of statistical analysis or modelling that requires interval-level measurements, because the purpose of applying a Rasch model is to obtain such measurements. Applications of Rasch models are described in a wide variety of sources, including Sivakumar, Durtis and Hungi (2005), Bezruzscko (2005), Bond and Fox (2007), Fisher and Wright (1994), Masters and Keeves (1999), and the *Journal of Applied Measurement*.

Invariant Comparison and Sufficiency

The Rasch model for dichotomous data is often regarded as an item response theory (IRT) model with one item parameter. However, rather than being a particular IRT model, proponents of the model regard it as a model that possesses a property which distinguishes it from other IRT models. Specifically, the defining property of Rasch models is their formal or mathematical *embodiment* of the principle of invariant comparison. Rasch summarised the principle of invariant comparison as follows:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion (Rasch, 1961, p. 332). Rasch models embody this principle because their formal structure permits algebraic separation of the person and item parameters, in the sense that the person parameter can be *eliminated* during the process of statistical estimation of item parameters. This result is achieved through the use of conditional maximum likelihood estimation, in which the response space is partitioned according to person total scores. The consequence is that the raw score for an item or person is the sufficient statistic for the item or person parameter. That is to say, the person total score contains all information available within the specified context about the individual, and the item total score contains all information with respect to item, with regard to the relevant latent trait. The Rasch model requires a specific structure in the response data, namely a probabilistic Guttman structure.

In somewhat more familiar terms, Rasch models provide a basis and justification for obtaining person locations on a continuum from total scores on assessments. Although it is not uncommon to treat total scores directly as measurements, they are actually counts of discrete observations rather than measurements. Each observation represents the observable outcome of a comparison between a person and item. Such outcomes are directly analogous to the observation of the rotation of a balance scale in one direction or another. This observation would indicate that one or other object has a greater mass, but counts of such observations cannot be treated directly as measurements.

Rasch pointed out that the principle of invariant comparison is characteristic of measurement in physics using, by way of example, a two-way experimental frame of reference in which each instrument exerts a mechanical force upon solid bodies to produce acceleration. Rasch (1960/1980, pp. 112–3) stated of this context: “Generally: If for any two objects we find a certain ratio of their accelerations produced by one instrument, then the same ratio will be found for any other of the instruments”. It is readily shown that Newton’s second law entails that such ratios are inversely proportional to the ratios of the masses of the bodies.

The Mathematical form of the Rasch Model for Dichotomous Data

Let $X_{ni} = x \in \{0, 1\}$ be a dichotomous random variable where, for example, $x = 1$ denotes a correct response and $x = 0$ an incorrect response to a given assessment item. In the Rasch model for dichotomous data, the probability of the outcome $X_{ni} = 1$ is given by:

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}},$$

where β_n is the ability of person n and δ_i is the difficulty of item i . Thus, in the case of a dichotomous attainment item, $Pr \{X_{ni}=1\}$ is the probability of success upon interaction between the relevant person and assessment item. It is readily shown that the log odds, or logit, of correct response by a person to an item, based on the model, is equal to $\beta_n - \delta_i$. It can be shown that the log odds of a correct response by a person to one item, *conditional* on a correct response to one of two items, is equal to the difference between the item locations. For example,

$$\log\text{-odds}\{X_{n1} = 1 | r_n = 1\} = \delta_2 - \delta_1,$$

where r_n is the total score of person n over the two items, which implies a correct response to one or other of the items. Hence, the conditional log odds does not involve the person parameter β_n , which can therefore be *eliminated* by conditioning on the total score $r_n=1$. That is, by partitioning the responses according to raw scores and calculating the log odds of a correct response, an estimate $\delta_2 - \delta_1$ is obtained without involvement of β_n . More generally, a number of item parameters can be estimated iteratively through application of a process such as Conditional Maximum Likelihood estimation. While more involved, the same fundamental principle applies in such estimations.

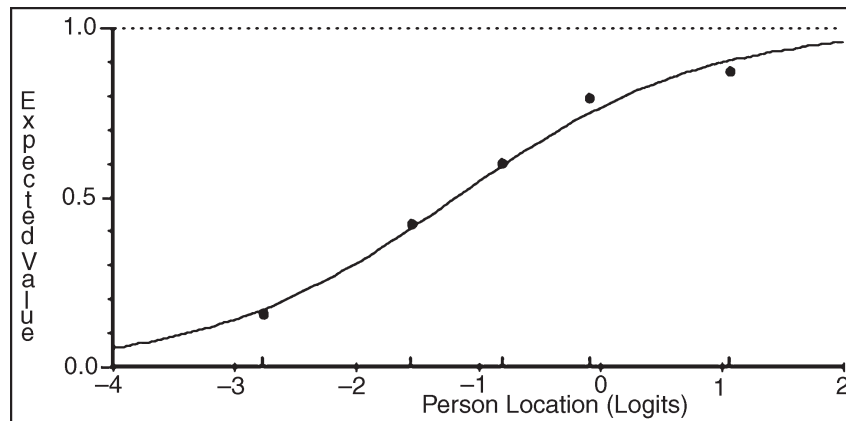


Fig. ICC for the Rasch Model Showing the Comparison Between Observed and Expected Proportions Correct for Five Class Intervals of Persons.

The ICC of the Rasch model for dichotomous data is shown in Figure above. The grey line maps a person with a location of approximately 0.2 on the latent continuum, to the probability of the discrete outcome $X_{ni}=1$ for items with different locations on the latent continuum. The location of an item is, by definition, that location at which the probability that $X_{ni}=1$ is equal to 0.5. The black circles represent the actual or observed proportions of persons within Class Intervals for which the outcome was observed.

For example, in the case of an assessment item used in the context of educational psychology, these could represent the proportions of persons who answered the item correctly. Persons are ordered by the estimates of their locations on the latent continuum and classified into Class Intervals on this basis in order to graphically

inspect the accordance of observations with the model. There is a close conformity of the data with the model. In addition to graphical inspection of data, a range of statistical tests of fit are used to evaluate whether departures of observations from the model can be attributed to random effects alone, as required, or whether there are systematic departures from the model.

The Polytomous form of the Rasch Model

The polytomous Rasch model, which is a generalisation of the dichotomous model, can be applied in contexts in which successive integer scores represent categories of increasing level or magnitude of a latent trait, such as increasing ability, motor function, endorsement of a statement, and so forth. The Polytomous response model is, for example, applicable to the use of Likert scales, grading in educational assessment, and scoring of performances by judges.

Other Considerations

A criticism of the Rasch model is that it is overly restrictive or prescriptive because it does not permit each item to have a different discrimination. A criticism specific to the use of multiple choice items in educational assessment is that there is no provision in the model for guessing because the left asymptote always approaches a zero probability in the Rasch model. These variations are available in models such as the two and three parameter logistic models (Birnbaum, 1968). However, the specification of uniform discrimination and zero left asymptote are necessary properties of the model in order to sustain sufficiency of the simple, unweighted raw score.

Verhelst and Glas (1995) derive Conditional Maximum Likelihood (CML) equations for a model they refer to as the One Parameter Logistic Model (OPLM). In algebraic form it appears to be identical with the 2PL model, but OPLM contains preset discrimination indexes rather than 2PL's estimated discrimination parameters. As noted by these authors, though, the problem one faces in estimation with estimated discrimination parameters is that the discriminations are unknown, meaning that the weighted raw score "is not a mere statistic, and hence it is impossible to use CML as an estimation method" (Verhelst and Glas, 1995, p. 217).

That is, sufficiency of the weighted "score" in the 2PL cannot be used according to the way in which a sufficient statistic is defined. If the weights are imputed instead of being estimated, as in OPLM, conditional estimation is possible and some of the properties of the Rasch model are retained (Verhelst, Glas and Verstralen, 1995; Verhelst and Glas, 1995). In OPLM, the values of the discrimination index are restricted to between 1 and 15. A limitation of this approach is that in practice, values of discrimination indexes must be preset as a starting point. This means some type of estimation of discrimination is involved when the purpose is to avoid doing so.

The Rasch model for dichotomous data inherently entails a single discrimination parameter which, as noted by Rasch (1960/1980, p. 121),

constitutes an arbitrary choice of the unit in terms of which magnitudes of the latent trait are expressed or estimated. However, the Rasch model requires that the discrimination is uniform across interactions between persons and items within a specified frame of reference (*i.e.*, the assessment context given conditions for assessment).

Application of the models provides diagnostic information regarding how well the criterion is met. Application of the models can also provide information about how well items or questions on assessments work to measure the ability or trait. Prominent advocates of Rasch models include Benjamin Drake Wright, David Andrich and Erling Andersen.

A comparison of Classical and Item Response Theories

Classical test theory (CTT) and IRT are largely concerned with the same problems but are different bodies of theory and entail different methods. Although the two paradigms are generally consistent and complementary, there are a number of points of difference:

- IRT makes stronger assumptions than CTT and in many cases provides correspondingly stronger findings; primarily, characterization of error. Of course, these results only hold when the assumptions of the IRT models are actually met.
- Although CTT results have allowed important practical results, the model-based nature of IRT affords many advantages over analogous CTT findings.
- CTT test scoring procedures have the advantage of being simple to compute (and to explain) whereas IRT scoring generally requires relatively complex estimation procedures.
- IRT provides several improvements in scaling items and people. The specifics depend upon the IRT model, but most models scale the difficulty of items and the ability of people on the same metric. Thus the difficulty of an item and the ability of a person can be meaningfully compared.
- Another improvement provided by IRT is that the parameters of IRT models are generally not sample- or test-dependent whereas true-score is defined in CTT in the context of a specific test. Thus IRT provides significantly greater flexibility in situations where different samples or test forms are used. These IRT findings are foundational for computerized adaptive testing.

It is worth also mentioning some specific similarities between CTT and IRT which help to understand the correspondence between concepts. First, Lord showed that under the assumption that θ is normally distributed, discrimination in the 2PL model is approximately a monotonic function of the point-biserial correlation. In particular:

$$a_i \cong \frac{\rho_{it}}{\sqrt{1 - \rho_{it}^2}}$$

where ρ_{it} is the point biserial correlation of item i . Thus, if the assumption holds, where there is a higher discrimination there will generally be a higher point-biserial correlation.

Another similarity is that while IRT provides for a standard error of each estimate and an information function, it is also possible to obtain an index for a test as a whole which is directly analogous to Cronbach's alpha, called the *separation index*. To do so, it is necessary to begin with a decomposition of an IRT estimate into a true location and error, analogous to decomposition of an observed score into a true score and error in CTT. Let

$$\hat{\theta} = \theta + \varepsilon$$

where θ is the true location, and ε is the error association with an estimate. Then $SE(\theta)$ is an estimate of the standard deviation of ε for person with a given weighted score and the separation index is obtained as follows

$$R_{\theta} = \frac{\text{var}[\theta]}{\text{var}[\hat{\theta}]} = \frac{\text{var}[\hat{\theta}] - \text{var}[\varepsilon]}{\text{var}[\hat{\theta}]}$$

where the mean squared standard error of person estimate gives an estimate of the variance of the errors, ε_n , across persons. The standard errors are normally produced as a by-product of the estimation process. The separation index is typically very close in value to Cronbach's alpha.

IRT is sometimes called *strong true score theory* or *modern mental test theory* because it is a more recent body of theory and makes more explicit the hypotheses that are implicit within CTT.

Equating

Test equating traditionally refers to the statistical process of determining comparable scores on different forms of an exam. It can be accomplished using either classical test theory or item response theory.

In item response theory, *equating* is the process of equating the units and origins of two scales on which the abilities of students have been estimated from results on different tests. The process is analogous to equating degrees Fahrenheit with degrees Celsius by converting measurements from one scale to the other. The determination of comparable scores is a by-product of equating that results from equating the scales obtained from test results.

Why is Equating Necessary

Suppose that Dick and Jane both take a test to become licensed in a certain profession. Because the high stakes (you get to practice the profession if you pass the test) may create a temptation to cheat, the organisation that oversees the test creates two forms. If we know that Dick scored 60 per cent on form A and Jane score 70 per cent on form B, do we know for sure which one has a better grasp of the material? What if form A is composed of very difficult items, while form B is relatively easy? Equating analyses are performed to address this very issue, so that scores are as fair as possible.

Equating in Item Response Theory

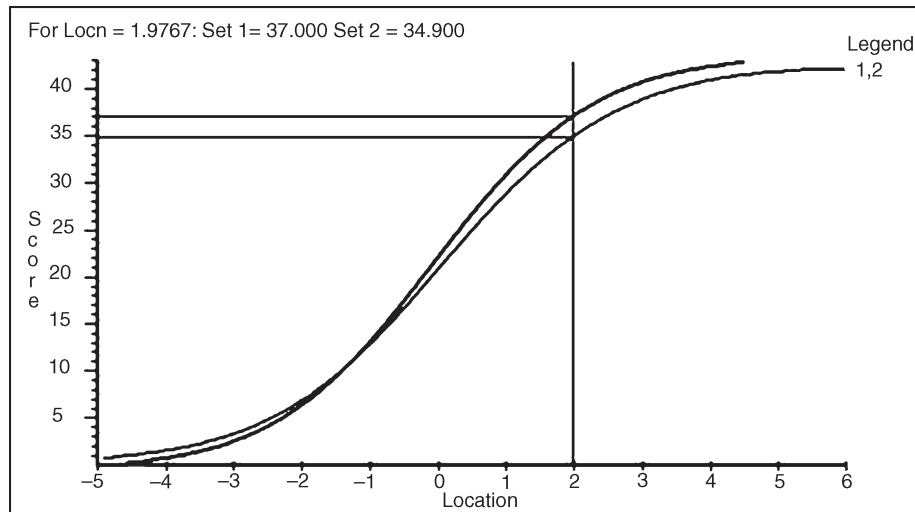


Fig. Test Characteristic Curves Showing the Relationship between Total Score and Person Location for two different Tests in Relation to a Common Scale. In this Example a total of 37 on Assessment 1 Equates to a Total of 34.9 on Assessment 2 as Shown by the Vertical Line.

In item response theory, person locations are estimated on a scale; *i.e.*, locations are estimated in relation to a unit and origin. It is common in educational assessment to employ tests in order to assess different groups of students with the intention of establishing a common scale by equating the origins, and sometimes units, of the scales obtained from response data from the different tests. The process is referred to as equating or test equating.

In item response theory, two different kinds of equating are horizontal and vertical equating. Vertical equating refers to the process of equating tests administered to groups of students with different abilities, such as students in different grades (years of schooling). Horizontal equating refers the equating of tests administered to groups with similar abilities; for example, two tests administered students in the same grade in two consecutive calendar years. Different tests are used to avoid practice effects. In terms of item response theory, equating is just a special case of the more general process of *scaling*, applicable when more than one test is used. In practice, though, scaling is often implemented separately for different tests and then the scales subsequently equated.

A distinction is often made between two methods of equating; *common person* and *common item* equating. Common person equating involves the administration of two tests to a common group of persons. The mean and standard deviation of the scale locations of the groups on the two tests are equated using a linear transformation. Common item equating involves the use of a set of common items referred to as the anchor test embedded in two different tests. The mean item location of the common items is equated.

Classical Approaches to Equating

In classical test theory, mean equating simply adjusts the distribution of scores so that the mean of one form is comparable to the mean of the other form. While mean equating is attractive because of its simplicity, it lacks flexibility, namely accounting for the possibility that the standard deviations of the forms differ.

Linear equating adjusts so that the two forms have a comparable mean and standard deviation. There are several types of linear equating that differ in the assumptions and mathematics used to estimate parameters. The Tucker and Levine Observed Score methods estimate the relationship between observed scores on the two forms, while the Levine True Score method estimates the relationship between true scores on the two forms. Equipercentile equating determines the equating relationship as one where a score could have an equivalent percentile on either form. This relationship can be nonlinear. Unlike with item response theory, equating based on classical test theory is somewhat distinct from scaling. Equating is a raw-to-raw transformation in that it estimates a raw score on Form B that is equivalent to each raw score on the base Form A. Any scaling transformation used is then applied on top of, or with, the equating.

EDUCATIONAL ASSESSMENT

Educational assessment is the process of documenting, usually in measurable terms, knowledge, skills, attitudes, and beliefs. Assessment can focus on the individual learner, the learning community (class, workshop, or other organised group of learners), the institution, or the educational system as a whole (also known as granularity). The final purposes and assessment practices in education depends on the *theoretical framework* of the practitioners and researchers, their assumptions and beliefs about the nature of human mind, the origin of knowledge and the process of learning.

Alternate Meanings

According to the Merriam-Webster online dictionary the word assessment comes from the root word assess which is defined as:

1. To determine the rate or amount of (as a tax)
2. To impose (as a tax) according to an established rate b: to subject to a tax, charge, or levy
3. To make an official valuation of (property) for the purposes of taxation
4. To determine the importance, size, or value of (assess a problem)
5. To charge (a player or team) with a foul or penalty

Assessment in education is best described as an action “to determine the importance, size, or value of.”

Types

The term assessment is generally used to refer to all activities teachers use to help students learn and to gauge student progress. Though the notion of assessment

is generally more complicated than the following categories suggest, assessment is often divided for the sake of convenience using the following distinctions:

1. Initial, formative, and summative
2. Objective and subjective
3. Referencing (criterion-referenced, norm-referenced, and ipsative)
4. Informal and formal.

Initial, Formative and Summative

Assessment is often divided into initial, formative, and summative categories for the purpose of considering different objectives for assessment practices:

- *Initial assessment*: Also referred to as pre-assessment or diagnostic assessment, initial assessments are conducted prior to instruction or intervention to establish a baseline from which individual student growth can be measured.
- *Formative assessment*: Formative assessment is generally carried out throughout a course or project. Formative assessment, also referred to as “educative assessment,” is used to aid learning. In an educational setting, formative assessment might be a teacher (or peer) or the learner, providing feedback on a student’s work and would not necessarily be used for grading purposes. Formative assessments can take the form of diagnostic, standardised tests.
- *Summative assessment*: Summative assessment is generally carried out at the end of a course or project. In an educational setting, summative assessments are typically used to assign students a course grade. Summative assessments are evaluative.

Educational researcher Robert Stake explains the difference between formative and summative assessment with the following analogy:

When the cook tastes the soup, that’s formative. When the guests taste the soup, that’s summative. Summative and formative assessment are often referred to in a learning context as *assessment of learning* and *assessment for learning* respectively. Assessment of learning is generally summative in nature and intended to measure learning outcomes and report those outcomes to students, parents and administrators. Assessment of learning generally occurs at the conclusion of a class, course, semester or academic year. Assessment for learning is generally formative in nature and is used by teachers to consider approaches to teaching and next steps for individual learners and the class.

A common form of formative assessment is *diagnostic assessment*. Diagnostic assessment measures a student’s current knowledge and skills for the purpose of identifying a suitable programme of learning. *Self-assessment* is a form of diagnostic assessment which involves students assessing themselves. *Forward-looking assessment* asks those being assessed to consider themselves in hypothetical future situations.

Performance-based assessment is similar to summative assessment, as it focuses on achievement. It is often aligned with the standards-based education

reform and outcomes-based education movement. Though ideally they are significantly different from a traditional multiple choice test, they are most commonly associated with standards-based assessment which use free-form responses to standard questions scored by human scorers on a standards-based scale, meeting, falling below or exceeding a performance standard rather than being ranked on a curve.

A well-defined task is identified and students are asked to create, produce or do something, often in settings that involve real-world application of knowledge and skills. Proficiency is demonstrated by providing an extended response. Performance formats are further differentiated into products and performances. The performance may result in a product, such as a painting, portfolio, paper or exhibition, or it may consist of a performance, such as a speech, athletic skill, musical recital or reading.

Objective and Subjective

Assessment (either summative or formative) is often categorized as either objective or subjective. Objective assessment is a form of questioning which has a single correct answer. Subjective assessment is a form of questioning which may have more than one correct answer (or more than one way of expressing the correct answer). There are various types of objective and subjective questions. Objective question types include true/false answers, multiple choice, multiple-response and matching questions. Subjective questions include extended-response questions and essays. Objective assessment is well suited to the increasingly popular computerized or online assessment format.

Some have argued that the distinction between objective and subjective assessments is neither useful nor accurate because, in reality, there is no such thing as “objective” assessment. In fact, all assessments are created with inherent biases built into decisions about relevant subject matter and content, as well as cultural (class, ethnic, and gender) biases.

Basis of Comparison

Test results can be compared against an established criterion, or against the performance of other students, or against previous performance:

- *Criterion-referenced assessment*, typically using a criterion-referenced test, as the name implies, occurs when candidates are measured against defined (and objective) criteria. Criterion-referenced assessment is often, but not always, used to establish a person’s competence (whether s/he can do something). The best known example of criterion-referenced assessment is the driving test, when learner drivers are measured against a range of explicit criteria (such as “Not endangering other road users”).
- *Norm-referenced assessment* (colloquially known as “grading on the curve”), typically using a norm-referenced test, is not measured against defined criteria. This type of assessment is relative to the student body

undertaking the assessment. It is effectively a way of comparing students. The IQ test is the best known example of norm-referenced assessment. Many entrance tests (to prestigious schools or universities) are norm-referenced, permitting a fixed proportion of students to pass (“passing” in this context means being accepted into the school or university rather than an explicit level of ability). This means that standards may vary from year to year, depending on the quality of the cohort; criterion-referenced assessment does not vary from year to year (unless the criteria change).

- *Ipsative assessment* is self comparison either in the same domain over time, or comparative to other domains within the same student.

Informal and Formal

Assessment can be either *formal* or *informal*. Formal assessment usually implies a written document, such as a test, quiz, or paper. A formal assessment is given a numerical score or grade based on student performance, whereas an informal assessment does not contribute to a student’s final grade such as this copy and pasted discussion question. An informal assessment usually occurs in a more casual manner and may include observation, inventories, checklists, rating scales, rubrics, performance and portfolio assessments, participation, peer and self-evaluation, and discussion.

Internal and External

Internal assessment is set and marked by the school (*i.e.*, teachers). Students get the mark and feedback regarding the assessment. External assessment is set by the governing body, and is marked by non-biased personnel. Some external assessments give much more limited feedback in their marking. However, in tests such as Australia’s Naplan, the criterion addressed by students is given detailed feedback in order for their teachers to address and compare the student’s learning achievements and also to plan for the future.

Standards of Quality

In general, high-quality assessments are considered those with a high level of reliability and validity. Approaches to reliability and validity vary, however.

Reliability: Reliability relates to the consistency of an assessment. A reliable assessment is one which consistently achieves the same results with the same (or similar) cohort of students. Various factors affect reliability—including ambiguous questions, too many options within a question paper, vague marking instructions and poorly trained markers.

Traditionally, the reliability of an assessment is based on the following:

1. *Temporal stability:* Performance on a test is comparable on two or more separate occasions.
2. *Form equivalence:* Performance among examinees is equivalent on different forms of a test based on the same content.

3. *Internal consistency*: Responses on a test are consistent across questions. For example: In a survey that asks respondents to rate attitudes towards technology, consistency would be expected in responses to the following questions:
 - “I feel very negative about computers in general.”
 - “I enjoy using computers.”

Reliability can also be expressed in mathematical terms as: $R_x = V_t/V_x$ where R_x is the reliability in the observed (test) score, X ; V_t and V_x are the variability in ‘true’ (*i.e.*, candidate’s innate performance) and measured test scores respectively. The R_x can range from 0 (completely unreliable), to 1 (completely reliable). An R_x of 1 is rarely achieved, and an R_x of 0.8 is generally considered reliable.

Validity

A valid assessment is one which measures what it is intended to measure. For example, it would not be valid to assess driving skills through a written test alone. A more valid way of assessing driving skills would be through a combination of tests that help determine what a driver knows, such as through a written test of driving knowledge, and what a driver is able to do, such as through a performance assessment of actual driving. Teachers frequently complain that some examinations do not properly assess the syllabus upon which the examination is based; they are, effectively, questioning the validity of the exam.

Validity of an assessment is generally gauged through examination of evidence in the following categories:

1. *Content*: Does the content of the test measure stated objectives?
2. *Criterion*: Do scores correlate to an outside reference? (ex: Do high scores on a 4th grade reading test accurately predict reading skill in future grades?)
3. *Construct*: Does the assessment correspond to other significant variables? (ex: Do ESL students consistently perform differently on a writing exam than native English speakers?)
4. *Face*: Does the item or theory make sense, and is it seemingly correct to the expert reader?

A good assessment has both validity and reliability, plus the other quality attributes noted above for a specific context and purpose. In practice, an assessment is rarely totally valid or totally reliable. A ruler which is marked wrong will always give the same (wrong) measurements. It is very reliable, but not very valid. Asking random individuals to tell the time without looking at a clock or watch is sometimes used as an example of an assessment which is valid, but not reliable. The answers will vary between individuals, but the average answer is probably close to the actual time. In many fields, such as medical research, educational testing, and psychology, there will often be a trade-off between reliability and validity. A history test written for high validity will have many essay and fill-in-the-blank questions.

It will be a good measure of mastery of the subject, but difficult to score completely accurately. A history test written for high reliability will be entirely multiple choice. It isn't as good at measuring knowledge of history, but can easily be scored with great precision. We may generalize from this.

The more reliable our estimate is of what we purport to measure, the less certain we are that we are actually measuring that aspect of attainment. It is also important to note that there are at least thirteen sources of invalidity, which can be estimated for individual students in test situations. They never are. Perhaps this is because their social purpose demands the absence of any error, and validity errors are usually so high that they would destabilize the whole assessment industry.

It is well to distinguish between "subject-matter" validity and "predictive" validity. The former, used widely in education, predicts the score a student would get on a similar test but with different questions. The latter, used widely in the workplace, predicts performance. Thus, a subject-matter-valid test of knowledge of driving rules is appropriate while a predictively valid test would assess whether the potential driver could follow those rules.

2

Educational Measurement

Educational measurement refers to the use of educational assessments and the analysis of data such as scores obtained from educational assessments to infer the abilities and proficiencies of students. The approaches overlap with those in psychometrics.

Overview

The aim of theory and practice in educational measurement is typically to measure abilities and levels of attainment by students in areas such as reading, writing, mathematics, science and so forth. Traditionally, attention focuses on whether assessments are reliable and valid. In practice, educational measurement is largely concerned with the analysis of data from educational assessments or tests. Typically, this means using total scores on assessments, whether they are multiple choice or open-ended and marked using marking rubrics or guides.

In technical terms, the pattern of scores by individual students to individual items is used to infer so-called scale locations of students, the “measurements”. This process is one form of scaling. Essentially, higher total scores give higher scale locations, consistent with the traditional and everyday use of total scores. If certain theory is used, though, there is not a strict correspondence between the ordering of total scores and the ordering of scale locations. The Rasch model provides a strict correspondence provided all students attempt the same test items, or their performances are marked using the same marking rubrics.

In terms of the broad body of purely mathematical theory drawn on, there is substantial overlap between educational measurement and psychometrics. However, certain approaches considered to be a part of psychometrics, including

Classical test theory, Item Response Theory and the Rasch model, were originally developed more specifically for the analysis of data from educational assessments.

One of the aims of applying theory and techniques in educational measurement is to try to place the results of different tests administered to different groups of students on a single or common scale through processes known as test equating. The rationale is that because different assessments usually have different difficulties, the total scores cannot be directly compared. The aim of trying to place results on a common scale is to allow comparison of the scale locations inferred from the totals via scaling processes.

OVERVIEW OF MEASUREMENT

Measurement is the process of determining the level of performance. This module presents basic ideas for obtaining valid, reliable, and efficient measurements, and illustrates how these are central to proper assessment, evaluation, and research.

Seminal Concepts

Measurement is important because people care about quality. *Quality* describes how good something is in the context of meeting human needs. Quality is a holistic combination of the inherent or distinctive attributes of a person, product, process, organisation, *etc.* Some examples of quality in higher-education contexts include

Quality of Knowledge

In a specific knowledge area (*e.g.*, hydrology, statistics, western history), quality involves an individual's depth, breath, and connections in the context of ideas and facts that comprise the knowledge area.

Quality of Performance

In a specific performance area (*e.g.*, teamwork, running a project, playing an oboe), quality describes how good the performance is.

Quality of a Product

For a given product (*e.g.*, technical report or journal paper, an original song), quality describes how good the product is.

Quality of an Organisation

For an academic unit (department, math tutoring centre, *etc.*), quality describes how effectively this unit meets the needs of key stakeholders. For a university, quality describes how effectively it meets the needs of the students.

Quality occurs on a scale that spans from low to medium to high to exceptional. *Measurement* is the process of assigning a number or qualitative

scale to indicate level of quality. Tools for making measurements have varied forms and names. Some common labels are *scoring guides*, *rubrics*, and *measures*. Here, we use the label measure to mean any tool that is used for the purposes of making a measurement.

Validity refers to how well the measurement process actually measures what it claims to measure. For example, a measurement of student writing should indicate the quality of the writing, and should not be influenced by things such as how much writing the student has done or whether or not the student has done things the way the teacher wanted them to be done.

Reliability refers to the repeatability of a measurement. That is, the more reliable a measurement, the more likely it is that the measurer will arrive at the same number or qualitative score if the measurement is repeated. In general, before the validity of a measurement process can be established, its reliability must be established. When multiple people use a measurement process, the level of consistency in their judgements is termed *inter-rater reliability*.

Quality in learning, assessment, evaluation, and research is enhanced by quality in measurement.

Rationale for Measurement

Assessment, evaluation, and research are three important processes in higher education. Although each is different, all three of them involve measurement.

Measurement targets should be meaningful to three different audiences: students, practitioners in the field, and researchers. Students respond best to explicit learning targets that involve authentic challenges connected with knowledge mastery, reasoning proficiency, product realisation, and professional expectations (Stiggins, 1996). Practitioners expect to see course outcomes that support the diverse roles within the discipline or profession and in the workplace. Researchers depend on a clearly conceptualised cognitive model that reflects the latest understanding of how learners represent knowledge and develop expertise in the domain (Pellegrino, Chudowsky, and Glaser, 2001). Researchers also expect alignment among the cognitive model, the methods used to observe performance, and the protocol for interpreting results. Educators vary both in their motivation for collecting data and in their skill in interpreting and reporting it. It is important to address the challenge of serving all three audiences with learning and growth that can be validly measured. The following sections explore the varying uses of measurement.

Role in Assessment

Assessment is a process of measuring and analysing a performance, a work product, or a learning skill to provide high-quality, timely feedback that gives assesseees clear and meaningful directives and insights to help them improve their future performance. Before a performance can be measured for assessment purposes, the criteria must be clearly defined and expectations or measures of each criterion must be set. The measurer will find it easier to provide specific

feedback that will be effective for strengthening future performance if he or she narrows the focus to three to five performance expectations. If the goal is to “grow” a performance, a work product, or a learning skill, assessment must occur early (and often) to allow students ample time to refine and improve. For example, if a central course outcome is to improve student writing, it will be important for instructors to conduct multiple “formative” measurements of performance on steps in the process of preparing a research paper. In this case, an instructor might use an analytic writing rubric for a research paper as the assessment tool to measure and collect data that provides feedback to the student. At intermediate times throughout the semester, instructors can measure specific performance expectations, providing both student and instructor with assessment data that can strengthen writing quality.

ROLE IN EVALUATION

Evaluation is the process of measuring the quality of a performance (*e.g.*, a work product or the use of a process) to make a judgement or to determine whether, or to what level, standards have been met. Evaluation is used in many academic arenas, such as graded assignments and exams, grade point average (GPA), promotion and tenure, or grant acquisition. Measurements that are used to make judgements are often based on external standards (*e.g.*, accrediting standards, agency policies, accountability for funding). Before any performance can be measured for evaluation purposes, the performance expectations (standards based on the measure) must be clear for each criterion of quality. Furthermore, the evaluation should be unbiased and be documented in a permanent record (*e.g.*, transcript, personnel file, grant record). In the case of a research paper, the final grade may be assigned using information from a score sheet associated with a writing rubric. The more a measurement tool requires an evaluator to explain his or her judgements about whether standards have been met, the less effective that measurement tool is for evaluation.

Role in Research

The purpose of measurement in research is to validate new knowledge within or across disciplines. Researchers begin with questions about a void in the existing body of knowledge; they then form hypotheses regarding relationships of measurable variables. Theory should be used to frame research questions and to guide methods for collecting reliable and valid data. In research, measurement falls into two categories: descriptive and experimental. If the researcher is attempting to answer a question descriptively, the appropriate tools include surveys, interviews or focus groups, conversational analysis, observation, ethnographies, or meta-analysis. If the researcher’s study is experimental in nature, the proper methods include randomised controlled trials, matched groups, baseline data, post-testing, and longitudinal designs. Each of these research designs or techniques requires certain kinds of measures that will result in data that can be appropriately analysed to provide a basis for interpretation (National

Research Council, 2002). Inferences drawn from the measurement should directly relate the evidence obtained to the hypothesis being investigated. The quality of a measure is very important because limitations, biases, and alternative interpretations will affect validity. Researchers want to know whether their findings can be generalised to a broader population or to multiple settings. The consistency of the measurement and the validity of the data are evidenced by the ability of other researchers to replicate the results.

Peer review and publication of research are essential for disseminating new knowledge to other practitioners as well as to the public.

Performance Measurement

Many educators are reluctant to apply measurement instruments and techniques to complex and integrated performances. Tasks like these are commonly referred to as *constructed-response outcomes*; they include learning portfolios, reflective journals, self-growth papers, capstone reports, project reports, and experiential narratives. Learning portfolios can include multiple *performance artifacts*, such as a sequence of art works produced during a course and accompanied by reflective journals and interpretive analyses. It is much easier to design constructed-response outcomes like portfolios than it is to create reliable and valid measures for assessing or evaluating their quality. To assess and/or to evaluate these complex outcomes, instructors often use custom-designed rubrics.

Educators' historic reluctance to adopt complex integrated performance outcomes stems in part from their assumptions about reliability and validity in measuring them. For many, selected-response instruments, such as multiple-choice and matching, are perceived to be more reliable and valid as well as easier to use. Instructors cannot measure performances that involve critical thinking, quality teaching, or service-learning projects by counting "correct" answers (Wiggins, 1998). These require qualitative judgements. As a result, some instructors opt to take advantage of the comfort that comes from using traditional select-response measurement instruments, and so spend most of their in-class time "covering the content" to align with "the test." But select-response tests are often not authentic measures of intended outcomes. For example, when one applies for a driver's license, the simple indicators of the driving test and written test do not represent and are not intended to represent all key driving performances.

A *competency* is a collection of knowledge, skills, and attitudes needed to perform a specific task effectively and efficiently at a defined level. A common question about a competency outcome is: What can the learner do at what level in a specific situation? *Movement* is documented growth in a transferable process or learning skill. A common question about a movement outcomes is: What does increased performance look like? *Accomplishments* are significant work products or performances that are externally valued or affirmed by an outside expert. A common question about an accomplishment outcome is: How well does student work compare with work products of practitioners in the field?

Experiences are interactions, emotions, responsibilities, and shared memories that clarify one's position in relation to oneself, a community, or discipline. A common question about an experience outcome is: How has this experience changed the learner? *Integrated performance* is the synthesis of prior knowledge, skills, processes, and attitudes with current learning needs to address a difficult challenge within a strict time frame and set of performance expectations. A common question about integrated performance is: How prepared are students to respond to a real-world challenge?

Over the last decade, rubrics have received considerable attention in education as tools for performance measurement. Rubrics provide explicit statements that describe different levels of performance and are worded in a way that covers the essence of what to look for when conducting qualitative measurements. Rubrics should reflect the best thinking about what constitutes good performance, a work product, or a learning skill. As discussed in *Fundamentals of Rubrics* (1.4.2), rubrics can be analytic (with an extensive set of factors and multiple scales) or holistic (with just a single scale). However, rubrics are only as robust as the clarity of purpose for measurement.

Concluding Thoughts

Measurement is foundational to classroom assessment, grading, programme evaluation, and educational research. In the physical sciences, quality measurement is a central event; in education, measurement involves a series of linked decisions that are more qualitative in nature. In both, the goal is to align outcomes, performance tasks, measurement methods, and data analysis. Educators in all disciplines must learn to apply their measurement skills to the multiple uses of measurement in education. Regardless of the discipline or profession, best practices include clear communication of purpose, well-selected targets for measurement, sound methods for data collection, and sampling to reduce bias and distortion. Faculty will become better teachers and researchers if they learn to seek consensus with their colleagues about what processes matter most in teaching and learning, and what tools measure learner growth most efficiently and effectively.

DEFINITION OF MEASUREMENT IN THE SOCIAL SCIENCES

The definition of measurement in the social sciences has a long history. A currently widespread definition, proposed by Stanley Smith Stevens (1946), is that measurement is “the assignment of numerals to objects or events according to some rule.” This definition was introduced in the paper in which Stevens proposed four levels of measurement. Although widely adopted, this definition differs in important respects from the more classical definition of measurement adopted in the physical sciences, which is that *measurement is the numerical estimation and expression of the magnitude of one quantity relative to another* (Michell, 1997).

Indeed, Stevens's definition of measurement was put forward in response to the British Ferguson Committee, whose chair, A. Ferguson, was a physicist. The committee was appointed in 1932 by the British Association for the Advancement of Science to investigate the possibility of quantitatively estimating sensory events. Although its chair and other members were physicists, the committee also included several psychologists. The committee's report highlighted the importance of the definition of measurement. While Stevens's response was to propose a new definition, which has had considerable influence in the field, this was by no means the only response to the report. Another, notably different, response was to accept the classical definition, as reflected in the following statement: Measurement in psychology and physics are in no sense different. Physicists can measure when they can find the operations by which they may meet the necessary criteria; psychologists have but to do the same. They need not worry about the mysterious differences between the meaning of measurement in the two sciences. These divergent responses are reflected in alternative approaches to measurement. For example, methods based on covariance matrices are typically employed on the premise that numbers, such as raw scores derived from assessments, are measurements. Such approaches implicitly entail Stevens's definition of measurement, which requires only that numbers are *assigned* according to some rule. The main research task, then, is generally considered to be the discovery of associations between scores, and of factors posited to underlie such associations.

On the other hand, when measurement models such as the Rasch model are employed, numbers are not assigned based on a rule. Instead, in keeping with Reese's statement above, specific criteria for measurement are stated, and the goal is to construct procedures or operations that provide data that meet the relevant criteria. Measurements are estimated based on the models, and tests are conducted to ascertain whether the relevant criteria have been met.

Instruments and Procedures

The first psychometric instruments were designed to measure the concept of intelligence. The best known historical approach involved the Stanford-Binet IQ test, developed originally by the French psychologist Alfred Binet. Intelligence tests are useful tools for various purposes. An alternative conception of intelligence is that cognitive capacities within individuals are a manifestation of a general component, or general intelligence factor, as well as cognitive capacity specific to a given domain.

Psychometrics is applied widely in educational assessment to measure abilities in domains such as reading, writing, and mathematics. The main approaches in applying tests in these domains have been Classical Test Theory and the more recent Item Response Theory and Rasch measurement models. These latter approaches permit joint scaling of persons and assessment items, which provides a basis for mapping of developmental continua by allowing descriptions of the skills displayed at various points along a continuum. Such

approaches provide powerful information regarding the nature of developmental growth within various domains. Another major focus in psychometrics has been on personality testing. There have been a range of theoretical approaches to conceptualising and measuring personality. Some of the better known instruments include the Minnesota Multiphasic Personality Inventory, the Five-Factor Model (or “Big 5”) and tools such as Personality and Preference Inventory and the Myers-Briggs Type Indicator. Attitudes have also been studied extensively using psychometric approaches. A common method in the measurement of attitudes is the use of the Likert scale. An alternative method involves the application of unfolding measurement models, the most general being the Hyperbolic Cosine Model.

Theoretical Approaches

Psychometricians have developed a number of different measurement theories. These include classical test theory (CTT) and item response theory (IRT). An approach which seems mathematically to be similar to IRT but also quite distinctive, in terms of its origins and features, is represented by the Rasch model for measurement. The development of the Rasch model, and the broader class of models to which it belongs, was explicitly founded on requirements of measurement in the physical sciences. Psychometricians have also developed methods for working with large matrices of correlations and covariance. Techniques in this general tradition include: factor analysis, a method of determining the underlying dimensions of data; multidimensional scaling, a method for finding a simple representation for data with a large number of latent dimensions; and data clustering, an approach to finding objects that are like each other. All these multivariate descriptive methods try to distil large amounts of data into simpler structures. More recently, structural equation modelling and path analysis represent more sophisticated approaches to working with large covariance matrices. These methods allow statistically sophisticated models to be fitted to data and tested to determine if they are adequate fits.

One of the main deficiencies in various factor analyses is a lack of consensus in cutting points for determining the number of latent factors. A usual procedure is to stop factoring when eigenvalues drop below one because the original sphere shrinks. The lack of the cutting points concerns other multivariate methods, also.

Key Concepts

Key concepts in classical test theory are reliability and validity. A reliable measure is one that measures a construct consistently across time, individuals, and situations. A valid measure is one that measures what it is intended to measure. Reliability is necessary, but not sufficient, for validity.

Both reliability and validity can be assessed statistically. Consistency over repeated measures of the same test can be assessed with the Pearson correlation coefficient, and is often called *test-retest reliability*. Similarly, the equivalence

of different versions of the same measure can be indexed by a Pearson correlation, and is called *equivalent forms reliability* or a similar term. Internal consistency, which addresses the homogeneity of a single test form, may be assessed by correlating performance on two halves of a test, which is termed *split-half reliability*; the value of this Pearson product-moment correlation coefficient for two half-tests is adjusted with the Spearman-Brown prediction formula to correspond to the correlation between two full-length tests. Perhaps the most commonly used index of reliability is Cronbach's α , which is equivalent to the mean of all possible split-half coefficients. Other approaches include the intra-class correlation, which is the ratio of variance of measurements of a given target to the variance of all targets.

There are a number of different forms of validity. Criterion-related validity can be assessed by correlating a measure with a criterion measure known to be valid.

When the criterion measure is collected at the same time as the measure being validated the goal is to establish *concurrent validity*; when the criterion is collected later the goal is to establish *predictive validity*. A measure has *construct validity* if it is related to measures of other constructs as required by theory. *Content validity* is a demonstration that the items of a test are drawn from the domain being measured. In a personnel selection example, test content is based on a defined statement or set of statements of knowledge, skill, ability, or other characteristics obtained from a *job analysis*.

Item response theory models the relationship between latent traits and responses to test items. Among other advantages, IRT provides a basis for obtaining an estimate of the location of a test-taker on a given latent trait as well as the standard error of measurement of that location.

For example, a university student's knowledge of history can be deduced from his or her score on a university test and then be compared reliably with a high school student's knowledge deduced from a less difficult test. Scores derived by classical test theory do not have this characteristic, and assessment of actual ability (rather than ability relative to other test-takers) must be assessed by comparing scores to those of a "norm group" randomly selected from the population. In fact, all measures derived from classical test theory are dependent on the sample tested, while, in principle, those derived from item response theory are not.

Standards of Quality

The considerations of validity and reliability typically are viewed as essential elements for determining the quality of any test. However, professional and practitioner associations frequently have placed these concerns within broader contexts when developing standards and making overall judgements about the quality of any test as a whole within a given context.

A consideration of concern in many applied research settings is whether or not the metric of a given psychological inventory is meaningful or arbitrary.

Testing Standards

In this field, the *Standards for Educational and Psychological Testing* place standards about validity and reliability, along with errors of measurement and related considerations under the general topic of test construction, evaluation and documentation. The second major topic covers standards related to fairness in testing, including fairness in testing and test use, the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities. The third and final major topic covers standards related to testing applications, including the responsibilities of test users, psychological testing and assessment, educational testing and assessment, testing in employment and credentialing, plus testing in programme evaluation and public policy.

0 Evaluation Standards

In the field of evaluation, and in particular educational evaluation, the Joint Committee on Standards for Educational Evaluation has published three sets of standards for evaluations. *The Personnel Evaluation Standards* was published in 1988, *The Programme Evaluation Standards* (2nd edition) was published in 1994, and *The Student Evaluation Standards* was published in 2003.

Each publication presents and elaborates a set of standards for use in a variety of educational settings. The standards provide guidelines for designing, implementing, assessing and improving the identified form of evaluation. Each of the standards has been placed in one of four fundamental categories to promote educational evaluations that are proper, useful, feasible, and accurate.

In these sets of standards, validity and reliability considerations are covered under the accuracy topic. For example, the student accuracy standards help ensure that student evaluations will provide sound, accurate, and credible information about student learning and performance.

Classical Test Theory

Classical test theory is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. Generally speaking, the aim of classical test theory is to understand and improve the reliability of psychological tests.

Classical test theory may be regarded as roughly synonymous with *true score theory*. The term “classical” refers not only to the chronology of these models but also contrasts with the more recent psychometric theories, generally referred to collectively as item response theory, which sometimes bear the appellation “modern” as in “modern latent trait theory”.

Classical test theory as we know it today was codified by Novick (1966) and described in classic texts such as Lord and Novick (1968) and Allen and Yen (1979/2002). The description of classical test theory below follows these seminal publications.

History

Classical Test Theory was born only after the following 3 achievements or ideas were conceptualised: one, a recognition of the presence of errors in measurements, two, a conception of that error as a random variable, and third, a conception of correlation and how to index it. In 1904, Charles Spearman was responsible for figuring out how to correct a correlation coefficient for attenuation due to measurement error and how to obtain the index of reliability needed in making the correction. Spearman's finding is thought to be the beginning of Classical Test Theory by some (Traub, 1997). Others who had an influence in the Classical Test Theory's framework include: George Udny Yule, Truman Lee Kelley, those involved in making the Kuder-Richardson Formulas, Louis Guttman, and, most recently, Melvin Novick, not to mention others over the next quarter century after Spearman's initial findings

Definitions

Classical test theory assumes that each person has a *true score*, T , that would be obtained if there were no errors in measurement. A person's true score is defined as the expected number-correct score over an infinite number of independent administrations of the test. Unfortunately, test users never observe a person's true score, only an *observed score*, X . It is assumed that *observed score* = *true score* plus some *error*:

$$X = T + E$$

observed score true score error

Classical test theory is concerned with the relations between the three variables X , T , and E in the population. These relations are used to say something about the quality of test scores. In this regard, the most important concept is that of *reliability*. The reliability of the observed test scores X , which is denoted as ρ^2_{XT} , is defined as the ratio of true score variance σ_T^2 to the observed score variance σ_X^2 :

$$\rho^2_{XT} = \frac{\sigma_T^2}{\sigma_X^2}$$

Because the variance of the observed scores can be shown to equal the sum of the variance of true scores and the variance of error scores, this is equivalent to

$$\rho^2_{XT} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

This equation, which formulates a signal-to-noise ratio, has intuitive appeal: The reliability of test scores becomes higher as the proportion of error variance in the test scores becomes lower and vice versa. The reliability is equal to the proportion of the variance in the test scores that we could explain if we knew the true scores. The square root of the reliability is the correlation between true and observed scores.

Evaluating Tests and Scores: Reliability

Reliability cannot be estimated directly since that would require one to know the true scores, which according to classical test theory is impossible. However, estimates of reliability can be obtained by various means. One way of estimating reliability is by constructing a so-called *parallel test*. The fundamental property of a parallel test is that it yields the same true score and the same observed score variance as the original test for every individual. If we have parallel tests x and x' , then this means that

$$\varepsilon(X_i) = \varepsilon(X'_i)$$

and

$$\sigma_{E_i}^2 = \sigma_{E'_i}^2$$

Under these assumptions, it follows that the correlation between parallel test scores is equal to reliability.

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2$$

Using parallel tests to estimate reliability is cumbersome because parallel tests are very hard to come by. In practice the method is rarely used. Instead, researchers use a measure of internal consistency known as Cronbach's α . Consider a test consisting of k items $u_j, j = 1, \dots, k$. The total test score is defined as the sum of the individual item scores, so that for individual i

$$X_i = \sum_{j=1}^k U_{ij}$$

Then Cronbach's alpha equals

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_j^k \sigma_{U_j}^2}{\sigma_X^2} \right)$$

Cronbach's α can be shown to provide a lower bound for reliability under rather mild assumptions. Thus, the reliability of test scores in a population is always higher than the value of Cronbach's α in that population. Thus, this method is empirically feasible and, as a result, it is very popular among researchers. Calculation of Cronbach's α is included in many standard statistical packages such as SPSS and SAS.

As has been noted above, the entire exercise of classical test theory is done to arrive at a suitable definition of reliability. Reliability is supposed to say something about the general quality of the test scores in question. The general idea is that, the higher reliability is, the better. Classical test theory does not say how high reliability is supposed to be. Too high a value for α , say over .9, indicates redundancy of items. Around .8 is recommended for personality research, while .9+ is desirable for individual high-stakes testing. These 'criteria' are not based on formal

arguments, but rather are the result of convention and professional practice. The extent to which they can be mapped to formal principles of statistical inference is unclear.

Evaluating Items: P and Item-Total Correlations

Reliability provides a convenient index of test quality in a single number, reliability. However, it does not provide any information for evaluating single items. Item analysis within the classical approach often relies on two statistics: the P-value (proportion) and the item-total correlation (point-biserial correlation coefficient). The P-value represents the proportion of examinees responding in the keyed direction, and is typically referred to as *item difficulty*. The item-total correlation provides an index of the discrimination or differentiating power of the item, and is typically referred to as *item discrimination*. In addition, these statistics are calculated for each response of the oft-used multiple choice item, which are used to evaluate items and diagnose possible issues, such as a confusing distractor. Such valuable analysis is provided by specially-designed psychometric software.

Alternatives

Classical test theory is an influential theory of test scores in the social sciences. In psychometrics, the theory has been superseded by the more sophisticated models in Item Response Theory (IRT) and Generalizability theory (G-theory). However, IRT is not included in standard statistical packages like SPSS and SAS, but there are IRT packages for the open source statistical programming language R (*e.g.*, CTT). While commercial packages routinely provide estimates of Cronbach's α , specialised psychometric software may be preferred for IRT or G-theory. However, general statistical packages often do not provide a complete classical analysis (Cronbach's α is only one of many important statistics), and in many cases, specialised software for classical analysis is also necessary.

Shortcomings of Classical Test Theory

One of the most important or well known shortcomings of Classical Test Theory is that examinee characteristics and test characteristics cannot be separated: each can only be interpreted in the context of the other. Another shortcoming lies in the definition of Reliability that exists in Classical Test Theory, which states that reliability is “the correlation between test scores on parallel forms of a test”.

The problem with this is that there are differing opinions of what parallel tests are. Various reliability coefficients provide either lower bound estimates of reliability or reliability estimates with unknown biases. A third shortcoming involves the standard error of measurement. The problem here is that, according to Classical Test Theory, the standard error of measurement is assumed to be the same for all examinees.

However, as Hambleton explains in his book, scores on any test are unequally precise measures for examinees of different ability, thus making the assumption of equal errors of measurement for all examinees implausible. A fourth, and final shortcoming of the Classical Test Theory is that it is test oriented, rather than item oriented. In other words, Classical Test Theory cannot help us make predictions of how well an individual or even a group of examinees might do on a test item.

Item Response Theory

In psychometrics, item response theory (IRT) also known as latent trait theory, strong true score theory, or modern mental test theory, is a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. Unlike simpler alternatives for creating scales as the simple sum questionnaire responses it does not assume that each item is equally difficult.

This distinguishes IRT from, for instance, the assumption in Likert scaling that “*All items are assumed to be replications of each other or in other words items are considered to be parallel instruments*” (p. 197). By contrast, item response theory treats the difficulty of each item (the ICCs) as information to be incorporated in scaling items.

It is based on the application of related mathematical models to testing data. Because it is generally regarded as superior to classical test theory, it is the preferred method for developing scales, especially when optimal decisions are demanded, as in so-called high-stakes tests *e.g.*, the Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT).

The name *item response theory* is due to the focus of the theory on the item, as opposed to the test-level focus of classical test theory. Thus IRT models the response of each examinee of a given ability to each item in the test. The term *item* is generic: covering all kinds of informative item. They might be multiple choice questions that have incorrect and correct responses, but are also commonly statements on questionnaires that allow respondents to indicate level of agreement (a rating or Likert scale), or patient symptoms scored as present/absent, or diagnostic information in complex systems.

IRT is based on the idea that the probability of a correct/keyed response to an item is a mathematical function of person and item parameters. The person parameter is construed as (usually) a single latent trait or dimension. Examples include general intelligence or the strength of an attitude. Parameters on which items are characterised include their difficulty (known as “location” for their location on the difficulty range), discrimination (slope or correlation) representing how steeply the rate of success of individuals varies with their ability, and a pseudo guessing parameter, characterising the (lower) asymptote at which even the least able persons will score due to guessing (for instance, 25 per cent for pure chance on a 4-item multiple choice item).

Overview

The concept of the item response function was around before 1950. The pioneering work of IRT as a theory occurred during the 1950s and 1960s. Three of the pioneers were the Educational Testing Service psychometrician Frederic M. Lord, the Danish mathematician Georg Rasch, and Austrian sociologist Paul Lazarsfeld, who pursued parallel research independently. Key figures who furthered the progress of IRT include Benjamin Drake Wright and David Andrich. IRT did not become widely used until the late 1970s and 1980s, when personal computers gave many researchers access to the computing power necessary for IRT.

Among other things, the purpose of IRT is to provide a framework for evaluating how well assessments work, and how well individual items on assessments work. The most common application of IRT is in education, where psychometricians use it for developing and refining exams, maintaining banks of items for exams, and equating for the difficulties of successive versions of exams (for example, to allow comparisons between results over time).

IRT models are often referred to as *latent trait models*. The term *latent* is used to emphasize that discrete item responses are taken to be *observable manifestations* of hypothesised traits, constructs, or attributes, not directly observed, but which must be inferred from the manifest responses. Latent trait models were developed in the field of sociology, but are virtually identical to IRT models.

IRT is generally regarded as an improvement over classical test theory (CTT). For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information. Some applications, such as computerised adaptive testing, are enabled by IRT and cannot reasonably be performed using only classical test theory. Another advantage of IRT over CTT is that the more sophisticated information IRT provides allows a researcher to improve the reliability of an assessment.

IRT entails three assumptions:

1. A unidimensional trait denoted by θ ;
2. Local independence of items;
3. The response of a person to an item can be modelled by a mathematical *item response function* (IRF).

The trait is further assumed to be measurable on a scale (the mere existence of a test assumes this), typically set to a standard scale with a mean of 0.0 and a standard deviation of 1.0. 'Local independence' means that items are not related except for the fact that they measure the same trait, which is equivalent to the assumption of unidimensionality, but presented separately because multidimensionality can be caused by other issues.

The topic of dimensionality is often investigated with factor analysis, while the IRF is the basic building block of IRT and is the centre of much of the research and literature.

The Item Response Function

The IRF gives the probability that a person with a given ability level will answer correctly. Persons with lower ability have less of a chance, while persons with high ability are very likely to answer correctly; for example, students with higher math ability are more likely to get a math item correct. The exact value of the probability depends, in addition to ability, on a set of *item parameters* for the IRF.

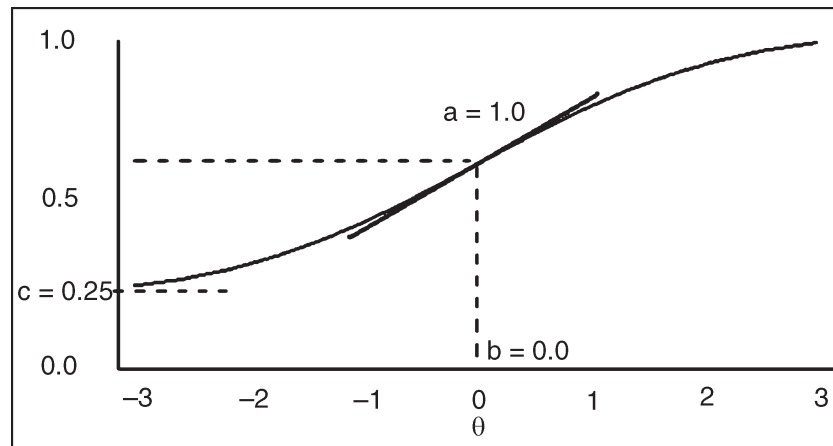


Fig. Three Parameter Logistic Model.

For example, in the *three parameter logistic* (3PL) model, the probability of a correct response to an item i is:

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}$$

where θ is the person (ability) parameter and a_i , b_i , and c_i are the item parameters. The item parameters simply determine the shape of the IRF and in some cases have a direct interpretation. The figure to the right depicts an example of the 3PL model of the ICC with an overlaid conceptual explanation of the parameters.

The item parameters can be interpreted as changing the shape of the standard logistic function:

$$P(t) = \frac{1}{1 + e^{-t}}$$

In brief, the parameters are interpreted as follows (dropping subscripts for legibility); b is most basic, hence listed first:

- b – difficulty, item location: $p(b) = (1 + c) / 2$, the half-way point between c_i (min) and 1 (max), also where the slope is maximized.
- a – discrimination, scale, slope: the maximum slope $p'(b) = a \cdot (1 - c) / 4$.
- c – pseudo-guessing, chance, asymptotic minimum $p(-\infty) = c$.

If $c = 0$ then these simplify to $p(b) = 1/2$ and $p'(b) = a/4$ meaning that b equals the 50 per cent success level (difficulty), and a (divided by four) is the maximum slope (discrimination), which occurs at the 50 per cent success level.

Further, the logit (log odds) of a correct response is $a(\theta - b)$ (assuming $c = 0$): in particular if ability θ equals difficulty b , there are even odds (1:1, so logit 0) of a correct answer, the greater the ability is above (or below) the difficulty the more (or less) likely a correct response, with discrimination a determining how rapidly the odds increase or decrease with ability.

In words, the standard logistic function has an asymptotic minimum of 0 ($c = 0$), is centred around 0 ($b = 0$, $P(0) = 1/2$), and has maximum slope $P'(0) = 1/4$. The a parameter stretches the horizontal scale, the b parameter shifts the horizontal scale, and the c compresses the vertical scale from $[0,1]$ to $[c,1]$. This is elaborated below. The parameter b_i represents the item location which, in the case of attainment testing, is referred to as the item difficulty. It is the point on θ where the IRF has its maximum slope, and where the value is half-way between the minimum value of c_i and the maximum value of 1. The example item is of medium difficulty since $b_i = 0.0$, which is near the centre of the distribution. Note that this model scales the item's difficulty and the person's trait onto the same continuum. Thus, it is valid to talk about an item being about as hard as Person A's trait level or of a person's trait level being about the same as Item Y's difficulty, in the sense that successful performance of the task involved with an item reflects a specific level of ability. The item parameter a_i represents the discrimination of the item: that is, the degree to which the item discriminates between persons in different regions on the latent continuum. This parameter characterises the slope of the IRF where the slope is at its maximum. The example item has $a_i = 1.0$, which discriminates fairly well; persons with low ability do indeed have a much smaller chance of correctly responding than persons of higher ability. For items such as multiple choice items, the parameter c_i is used in attempt to account for the effects of guessing on the probability of a correct response. It indicates the probability that very low ability individuals will get this item correct by chance, mathematically represented as a lower asymptote. A four-option multiple choice item might have an IRF like the example item; there is a 1/4 chance of an extremely low ability candidate guessing the correct answer, so the c_i would be approximately 0.25. This approach assumes that all options are equally plausible, because if one option made no sense, even the lowest ability person would be able to discard it, so IRT parameter estimation methods take this into account and estimate a c_i based on the observed data.

IRT Models

Broadly speaking, IRT models can be divided into two families: unidimensional and multidimensional. Unidimensional models require a single trait (ability) dimension θ . Multidimensional IRT models model response data hypothesised to arise from multiple traits. However, because of the greatly increased complexity, the majority of IRT research and applications utilise a unidimensional model. IRT models can also be categorised based on the number of scored responses. The typical multiple choice item is *dichotomous*; even though there may be four or five options, it is still scored only as correct/incorrect

(right/wrong). Another class of models apply to *polytomous* outcomes, where each response has a different score value. A common example of this are Likert-type items, *e.g.*, “Rate on a scale of 1 to 5.”

Number of IRT Parameters

Dichotomous IRT models are described by the number of parameters they make use of. The 3PL is named so because it employs three item parameters. The two-parameter model (2PL) assumes that the data have no guessing, but that items can vary in terms of location (b_i) and discrimination (a_i). The one-parameter model (1PL) assumes that guessing is a part of the ability and that all items that fit the model have equivalent discriminations, so that items are only described by a single parameter (b_i). This results in one-parameter models having the property of specific objectivity, meaning that the rank of the item difficulty is the same for all respondents independent of ability, and that the rank of the person ability is the same for items independently of difficulty. Thus, 1 parameter models are sample independent, a property that does not hold for two-parameter and three-parameter models. Additionally, there is theoretically a four-parameter model (4PL), with an upper asymptote, denoted by d_i where $1-c_i$ in the 3PL is replaced by d_i-c_i . However, this is rarely used. Note that the alphabetical order of the item parameters does not match their practical or psychometric importance; the location/difficulty (b_i) parameter is clearly most important because it is included in all three models. The 1PL uses only b_i , the 2PL uses b_i and a_i , the 3PL adds c_i , and the 4PL adds d_i .

The 2PL is equivalent to the 3PL model with $c_i = 0$, and is appropriate for testing items where guessing the correct answer is highly unlikely, such as fill-in-the-blank items (“What is the square root of 121?”), or where the concept of guessing does not apply, such as personality, attitude, or interest items (*e.g.*, “I like Broadway musicals. Agree/Disagree”). The 1PL assumes not only that guessing is not present (or irrelevant), but that all items are equivalent in terms of discrimination, analogous to a common factor analysis with identical loadings for all items. Individual items or individuals might have secondary factors but these are assumed to be mutually independent and collectively orthogonal.

Logistic and Normal IRT Models

An alternative formulation constructs IRFs based on the normal probability distribution; these are sometimes called *normal ogive models*. For example, the formula for a two-parameter normal-ogive IRF is:

$$p_i(\theta) = \Phi\left(\frac{\theta - b_i}{\sigma_i}\right)$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution.

The normal-ogive model derives from the assumption of normally distributed measurement error and is theoretically appealing on that basis. Here b_i is, again,

the difficulty parameter. The discrimination parameter is σ_i , the standard deviation of the measurement error for item i , and comparable to $1/a_i$. One can estimate a normal-ogive latent trait model by factor-analysing a matrix of tetrachoric correlations between items. This means it is technically possible to estimate a simple IRT model using general-purpose statistical software.

With rescaling of the ability parameter, it is possible to make the 2PL logistic model closely approximate the cumulative normal ogive. Typically, the 2PL logistic and normal-ogive IRFs differ in probability by no more than 0.01 across the range of the function. The difference is greatest in the distribution tails, however, which tend to have more influence on results.

The latent trait/IRT model was originally developed using normal ogives, but this was considered too computationally demanding for the computers at the time (1960s). The logistic model was proposed as a simpler alternative, and has enjoyed wide use since. More recently, however, it was demonstrated that, using standard polynomial approximations to the normal *cdf*, the normal-ogive model is no more computationally demanding than logistic models.

The Rasch Model

The Rasch model is often considered to be the 1PL IRT model. However, proponents of Rasch modelling prefer to view it as a completely different approach to conceptualising the relationship between data and the theory. Like other statistical modelling approaches, IRT emphasizes the primacy of the fit of a model to observed data, while the Rasch model emphasizes the primacy of the requirements for fundamental measurement, with adequate data-model fit being an important but secondary requirement to be met before a test or research instrument can be claimed to measure a trait. Operationally, this means that the IRT approaches include additional model parameters to reflect the patterns observed in the data (*e.g.*, allowing items to vary in their correlation with the latent trait), whereas the Rasch approach requires both the data fit the Rasch model and that test items and examinees confirm to the model, before claims regarding the presence of a latent trait can be considered valid.

Therefore, under Rasch models, misfitting responses require diagnosis of the reason for the misfit, and may be excluded from the data set if substantive explanations can be made that they do not address the latent trait. Thus, the Rasch approach can be seen to be a confirmatory approach, as opposed to exploratory approaches that attempt to model the observed data. As in any confirmatory analysis, care must be taken to avoid confirmation bias.

The presence or absence of a guessing or pseudo-chance parameter is a major and sometimes controversial distinction. The IRT approach includes a left asymptote parameter to account for guessing in multiple choice examinations, while the Rasch model does not because it is assumed that guessing adds randomly distributed noise to the data. As the noise is randomly distributed, it is assumed that, provided sufficient items are tested, the rank-ordering of persons along the latent trait by raw score will not change, but will simply undergo a

linear rescaling. Three-parameter IRT, by contrast, achieves data-model fit by selecting a model that fits the data, at the expense of sacrificing specific objectivity. In practice, the Rasch model has at least two principal advantages in comparison to the IRT approach. The first advantage is the primacy of Rasch's specific requirements, which (when met) provides *fundamental* person-free measurement (where persons and items can be mapped onto the same invariant scale). Another advantage of the Rasch approach is that estimation of parameters is more straightforward in Rasch models due to the presence of sufficient statistics, which in this application means a one-to-one mapping of raw number-correct scores to Rasch θ estimates.

Analysis of Model Fit

As with any use of mathematical models, it is important to assess the fit of the data to the model. If item misfit with any model is diagnosed as due to poor item quality, for example confusing distractors in a multiple-choice test, then the items may be removed from that test form and rewritten or replaced in future test forms. If, however, a large number of misfitting items occur with no apparent reason for the misfit, the construct validity of the test will need to be reconsidered and the test specifications may need to be rewritten. Thus, misfit provides invaluable diagnostic tools for test developers, allowing the hypotheses upon which test specifications are based to be empirically tested against data.

There are several methods for assessing fit, such as a chi-square statistic, or a standardised version of it. Two and three-parameter IRT models adjust item discrimination, ensuring improved data-model fit, so fit statistics lack the confirmatory diagnostic value found in one-parameter models, where the idealised model is specified in advance. Data should not be removed on the basis of misfitting the model, but rather because a construct relevant reason for the misfit has been diagnosed, such as a non-native speaker of English taking a science test written in English. Such a candidate can be argued to not belong to the same population of persons depending on the dimensionality of the test, and, although one parameter IRT measures are argued to be sample-independent, they are not population independent, so misfit such as this is construct relevant and does not invalidate the test or the model. Such an approach is an essential tool in instrument validation. In two and three-parameter models, where the psychometric model is adjusted to fit the data, future administrations of the test must be checked for fit to the same model used in the initial validation in order to confirm the hypothesis that scores from each administration generalise to other administrations. If a different model is specified for each administration in order to achieve data-model fit, then a different latent trait is being measured and test scores cannot be argued to be comparable between administrations.

Information

One of the major contributions of item response theory is the extension of the concept of reliability. Traditionally, reliability refers to the precision of

measurement (*i.e.*, the degree to which measurement is free of error). And traditionally, it is measured using a single index defined in various ways, such as the ratio of true and observed score variance. This index is helpful in characterising a test's average reliability, for example in order to compare two tests. But IRT makes it clear that precision is not uniform across the entire range of test scores. Scores at the edges of the test's range, for example, generally have more error associated with them than scores closer to the middle of the range.

Item response theory advances the concept of item and test information to replace reliability. Information is also a *function* of the model parameters. For example, according to Fisher information theory, the item information supplied in the case of the 1PL for dichotomous response data is simply the probability of a correct response multiplied by the probability of an incorrect response, or,

$$I(\theta) = p_i(\theta)q_i(\theta).$$

The standard error of estimation (SE) is the reciprocal of the test information of at a given trait level, is the

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

Thus more information implies less error of measurement.

For other models, such as the two and three parameters models, the discrimination parameter plays an important role in the function. The item information function for the two parameter model is

$$I(\theta) = a_i^2 p_i(\theta)q_i(\theta).$$

The item information function for the three parameter model is

$$I(\theta) = a_i^2 \frac{(p_i(\theta) - c_i)^2}{(1 - c_i)^2} \frac{q_i(\theta)}{p_i(\theta)}$$

In general, item information functions tend to look bell-shaped. Highly discriminating items have tall, narrow information functions; they contribute greatly but over a narrow range. Less discriminating items provide less information but over a wider range. Plots of item information can be used to see how much information an item contributes and to what portion of the scale score range. Because of local independence, item information functions are additive. Thus, the test information function is simply the sum of the information functions of the items on the exam. Using this property with a large item bank, test information functions can be shaped to control measurement error very precisely. Characterising the accuracy of test scores is perhaps the central issue in psychometric theory and is a chief difference between IRT and CTT. IRT findings reveal that the CTT concept of reliability is a simplification. In the place of reliability, IRT offers the test information function which shows the degree of precision at different values of theta, θ .

These results allow psychometricians to (potentially) carefully shape the level of reliability for different ranges of ability by including carefully chosen

items. For example, in a certification situation in which a test can only be passed or failed, where there is only a single “cutscore,” and where the actually passing score is unimportant, a very efficient test can be developed by selecting only items that have high information near the cutscore. These items generally correspond to items whose difficulty is about the same as that of the cutscore.

Scoring

The person parameter θ represents the magnitude of *latent trait* of the individual, which is the human capacity or attribute measured by the test. It might be a cognitive ability, physical ability, skill, knowledge, attitude, personality characteristic, *etc.* The estimate of the person parameter - the “score” on a test with IRT - is computed and interpreted in a very different manner as compared to traditional scores like number or percent correct. The individual’s total number-correct score is not the actual score, but is rather based on the IRFs, leading to a weighted score when the model contains item discrimination parameters. It is actually obtained by multiplying the item response function for each item to obtain a *likelihood function*, the highest point of which is the *maximum likelihood estimate* of θ . This highest point is typically estimated with IRT software using the Newton-Raphson method. While scoring is much more sophisticated with IRT, for most tests, the (linear) correlation between the theta estimate and a traditional score is very high; often it is .95 or more. A graph of IRT scores against traditional scores shows an ogive shape implying that the IRT estimates separate individuals at the borders of the range more than in the middle.

An important difference between CTT and IRT is the treatment of measurement error, indexed by the standard error of measurement. All tests, questionnaires, and inventories are imprecise tools; we can never know a person’s *true score*, but rather only have an estimate, the *observed score*. There is some amount of random error which may push the observed score higher or lower than the true score. CTT assumes that the amount of error is the same for each examinee, but IRT allows it to vary. Also, nothing about IRT refutes human development or improvement or assumes that a trait level is fixed. A person may learn skills, knowledge or even so called “test-taking skills” which may translate to a higher true-score. In fact, a portion of IRT research focuses on the measurement of change in trait level.

BLOOM’S TAXONOMY—EXPANDING ITS MEANING

Educational objectives indicate what students should attend to and put effort into learning; they are “explicit formulations of the ways in which students are expected to be changed by the educative process” (Bloom, 1956). Bloom’s taxonomy provides a well-accepted pedagogical framework for classifying vast numbers of educational objectives into useful structures. Benjamin Bloom’s pioneering work on learning was initiated in 1948, when he headed a team of

educators and psychologists investigating three major learning domains: cognitive, affective, and psychomotor. Over the last half-century, the theoretical framework produced by this team has facilitated analyses of learning objectives classification, criteria for performance-based learning, and levels of mastery in learning (Simon, 2000).

To the extent that the goal of education is the diffusion of knowledge through learning, a description of Bloom's taxonomy represents a seminal work in developing and implementing high quality instruction. There are six different levels in the cognitive domain of factual and conceptual knowledge progressing from elementary to complex.

Evolution of Bloom's Taxonomy

Over the past 40 years Bloom's work has been translated into more than twenty languages and has provided a basis for test design and curriculum development. Many modern interpretations of Bloom's taxonomy are found in the literature. Recently Anderson and Krathwohl (2001) expanded the single dimension of the original taxonomy into a two-dimensional framework consisting of factual/conceptual knowledge and cognitive processes.

High quality educational objectives combine both elements as seen in the following example: "The student will learn to distinguish (cognitive process) among confederate, federal, and unitary systems of government (knowledge)." Apple and Krumsieg (2001) clarified some of the definitions found in the original taxonomy by viewing it in terms of transferable knowledge that progresses in complexity through the six levels. The most basic level, using Apple and Krumsieg's labels, involves information (knowledge in Bloom), followed by knowledge (comprehension), knowledge skill (application), problem solution (analysis), new knowledge (synthesis), and finally evaluation (peer-reviewed knowledge). This model of learning has supported the development of a learning process methodology for efficiently and effectively advancing the level of student knowledge.

There is also extensive educational research aimed at moving beyond the cognitive domain in formal education by focusing more attention on the affective and psychomotor domains. Although this is not the focus of this module, it is important to be aware of these developments. Tinto (1993) and Shank (1994) have published significant works in this area, arguing that academics must change the way teaching is performed, by paying special attention to the intrapersonal and interpersonal contexts of learning. Tinto examined learning communities in depth, while Shank promotes the perspective that the only way learning occurs is "by doing."

Levels of Learner Knowledge

Bloom's taxonomy has been adapted and transformed by Apple and Krumsieg (2001). According to their Learning Process Methodology, five levels of learner knowledge are observable in college classrooms. Information acquisition

occupies the lowest level and is typified by memorisation of information. Conceptual understanding represents the next higher level and is the result of combining information elements to achieve understanding and meaning. Application is the ability to apply knowledge in a new context. Working expertise is the ability to understand the logical constructs and apply knowledge without expert prompting. Research is the goal of graduate study and is the ability to create novel discoveries from basic elements and logical constructs. The “evaluation level” in Bloom is considered separately as part of assessment, which can take place at any level.

Brookfield (1987) argues that learning is promoted by asking questions that challenge students’ understanding at the appropriate level. Good questions can also stimulate students’ curiosity and allow the teacher to probe current understanding as well as assess the effectiveness of past instructional activities. Inquiry as a learning method requires active participation both by the students and teachers. The combination of these two concepts creates a useful tool for teachers to use in classroom applications. It provides key words and questions that are appropriate to ask students at each level of learning and demonstrates the link with Bloom’s taxonomy of educational objectives.

Classroom Application

For example, introducing methodologies and studying their elements at Levels One and Two is a particularly effective way to accelerate the creation of transferable knowledge at Levels Three and Four.

Fundamental to all aspects of educational processes is the knowledge that results from experiencing applications of knowledge.

3

Measurement, Evaluation and Research

Ways of Knowing

In general, there are four ways or methods by which we can ascertain the truth of something. First, we can know something is true because we trust the source of the information. For example, we may read a textbook or review a research study. We can also use references other than scientific studies such as religious literature (*e.g.*, the Talmud, the Bible, the Koran, *etc.*) In both cases, the information has been revealed to us and we trust the source of the information. Second, we may know something is true through intuition or personal inspiration.

We may feel strongly that we have been “guided” to truth through an insight that is unique and personal. A third way of knowing is through personal experience. This is often a powerful approach to many people. A fourth way of knowing is through reason or thinking logically and critically about the first three.

Each of these ways of knowing is potentially flawed. We may read something from an otherwise credible source who has made a mistake relative to a particular issue. We may also have an inspiration that upon further investigation it may prove to be incorrect. The possibility of error through personal experience is well known by way of optical illusions. And obviously, reason is capable of error since a number of scientists have different explanations for the same set of data and teachers of religion have different explanations of the same inspired text.

Kerlinger (1973) summarizing the writings of the philosopher Charles Pierce (as cited in Buchler, 1955 and Cohen & Nagel, 1934), provides a slightly different

view of the four methods by which we determine truth. The first is the method of tenacity whereby truth is what is known to the individual or group. It simply is true.

The second is the method of authority in which truth is established through a trusted source such as God, tradition, or public sanction. The third is *a priori* method or the method of intuition. The fourth method is the scientific method which attempts to define a process for defining truth that produces results verifiable by others and is self-correcting. Kerlinger's definition of scientific research is that it is a "systematic, controlled, empirical, and critical investigation of hypothetical propositions about the presumed relations among natural phenomena" (p. 11).

Science, in terms of the ways of knowing discussed by Kerlinger (1973), might be considered a special case of the combination of experience and reason. While inspiration or intuition often plays an important role in scientific discovery, it must be subjected to experience that can be publicly verified and reason before it is accepted. The same holds true of revealed information; it is expected that we replicate or test out someone else's experience or ideas as reported in scientific or non-scientific literature or religious scripture.

Classification of Scientific Knowledge

Knowledge, especially scientific knowledge, can be classified into six categories:

1. *Facts—an idea or action that can be verified—Example:* Names and dates of important activities; population of the United States in the latest census.
2. *Concepts—Rules that allow for categorization of events, places, people ideas, etc.—Example:* A DESK is a piece of furniture (also a concept) designed with a flat top for writing; a CHAIR is a piece of furniture designed for sitting; a CHAIR with a flat surface attached to it that is designed for writing is also called a DESK.
3. *Principles—relationship(s) between/among facts and/or concepts—Example:* The number of children in the family is related to the average scores on nationally standardized achievement tests for those children.
4. *Hypotheses—educated guess about relationships (principles)—Example:* for lower-division, undergraduate students study habits is a better predictor of success in a college course than is a measure of intelligence or reading comprehension.
5. *Theories—set of facts, concepts, and principles that allow description and explanation—Example:* Piaget's theory of cognitive development, Erikson's theory of socioemotional development, Skinner's theory of operant conditioning; and.
6. *Laws—firmly established, thoroughly tested, principle or theory—Example:* A fixed interval schedule for delivering reinforcement produces a scalloping effect on behaviour.

The human mind does not think or reason in terms of discrete elements or “facts.” Rather it processes information in terms of concepts or the rules for categorizing facts. When we build relationships among facts and concepts (*i.e.*, develop principles), we are able to remember, understand, and access an astonishing amount of information. We are also able to make predictions from present to future circumstances. However, it is when we develop theories (add explanations to facts, concepts, and principles) and laws (empirically validate principles and theories) that we accomplish the highest goal of science—to control the variables we are studying.

The different types of scientific studies relate rather well to this classification of knowledge: if we desire to develop facts and concepts, descriptive studies can serve our purpose. If we desire to develop principles, we probably need to use correlational studies that will allow us to make predictions from present circumstances. Of course, hypotheses can be developed from facts and concepts and then verified by either correlational or experimental research. Theories are developed when one produces an explanation for the facts, concepts, or principles. Notice that extensive research support is not necessary to develop a theory. Laws can only be derived from experimental research. Unfortunately, solid experimental evidence is not widely generated in educational psychology and we have, therefore, not produced an extensive number of laws for teaching and learning.

Again, the controversy over what knowledge source to use in developing theories and laws comes into play. If we are simply material beings living in a material universe, then limiting our explanations to those derived from science is appropriate. However, if we are, in essence, spiritual beings connected to a spiritual aspect of the universe that is capable of influencing our material existence, then the exclusive use of science is unwarranted. Again, however, since this is a science course we will limit ourselves to knowledge derived using the scientific method, even though by doing so we may have omitted valuable information. It will be necessary for each individual to integrate this knowledge with that derived from experience, intuition, religion, and/or philosophy.

EDUCATIONAL RESEARCH, MEASUREMENT AND EVALUATION

The Educational Research, Measurement, and Evaluation (ERME) programme at the Lynch School combines the study of research design, statistical methods, and testing and assessment with a research focus on major contemporary education policy issues.

The programme is designed to prepare students for research and academic careers in education, social sciences and human services.

The ERME programme offers two degrees—the Master of Education (M.Ed.) and the Doctor of Philosophy (Ph.D.). The programme provides in-depth expertise in quantitative and experimental methods for research and evaluation, with opportunities for students to tailor coursework to their particular interests and background.

Courses cover three main areas:

- Research design and methods
- Statistical methods
- Testing and assessment

<i>Sequence</i>	<i>Course topics</i>
Research design and methods	<ul style="list-style-type: none"> • Interpretation and evaluation of research • Models of curriculum and programme evaluation • Design of quantitative research • Design of experiments • Survey research methods • Seminars in educational measurement and research
Statistical methods	<ul style="list-style-type: none"> • Introductory and intermediate statistics • General linear models • Multivariate statistics • Psychometric theory • Seminars in statistical and measurement topics
Testing and assessment	<ul style="list-style-type: none"> • Classroom assessment • Large scale assessment • Public policy • Seminar on current issues in testing and assessment • Practical in technology-enhanced assessment

Students may also conduct an independent study to pursue a particular area of interest. The ERME programme has been training Ph.D.'s to examine educational programmes, design quantitative research studies, develop assessment instruments, and analyze educational data to help inform policy-making for almost 40 years. Its outstanding faculty and rich intellectual resources provide students with the ideal setting for learning and professional growth. A survey of similar doctoral programmes in North America, presented in a paper at the annual meeting of the National Council on Measurement in Education, April 2001, placed the Lynch School programme in the top 10 overall.

ERME students have the opportunity to work on research projects with individual faculty members or in one of the Lynch School's research centres. More specifically, many of the students work as research assistants in the Centre for the Study of Testing, Evaluation, and Educational Policy (CSTEPP) and in the TIMSS and PIRLS International Study Centre (ISC). These projects often afford students with the opportunity to author and present papers at educational research conferences such as those held by the American Educational Research Association (AERA) and New England Educational Research Organization (NEERO).

Graduates are qualified for a wide range of positions in federal, state, and local government agencies; private research companies; non-profit organizations and foundations; and schools, colleges and universities.

At the doctoral level, opportunities for job placement include positions as:

- Principal investigators on research projects
- Policy researchers and analysts

- Programme evaluators
- Measurement specialists
- Methodological consultants
- College and university professors

At the master's level, job opportunities include mid-level research, testing, and evaluation positions in education and the social sciences; it also prepares students to pursue advanced graduate studies.

Sample of Career Descriptions

Master of Education (M.Ed.): The master's degree curriculum includes coursework in research design, statistics, classroom assessment, large scale data collection, programme evaluation, and education policy. A minimum of 30 credits and satisfactory performance on a comprehensive examination are required for the M.Ed. degree.

Programme of Study

Doctor of Philosophy (Ph.D.): The doctoral curriculum emphasizes research methodology and data analysis and includes advanced coursework in research design, statistical methods, and testing and assessment as well as seminars in statistical and measurement topics. The doctoral degree requires a minimum of 54 credits beyond the M.Ed. and satisfactory completion of comprehensive exams and a dissertation.

Practical Process Engineering for Higher Education

The Glendale Community College (GCC) and Oklahoma City Community College (OCC) process engineering programmes are working to increase the efficiency of college processes, thereby reducing the overhead cost per student, while providing higher quality services. GCC's project, PEP, applies the discipline of systems engineering to processes creating a highly structured methodology; employs a pro-active change management programme; uses existing college resources; and integrates the programme into planning and management activities. OCC's Project VIRGIL, applies a different strategy, but also utilizes a highly-structured design model. This model is influenced by cross-functional teams with limited outside resources to support the process engineering endeavours.

Introduction: Post-secondary education is now facing numerous challenges. Many external sources are placing pressures on colleges and universities. Enrolment are rising at the same time that colleges are experiencing increased competition for limited dollars.

This is especially true for community colleges with new programmes such as Welfare to Work. Also, the *Chronicle of Higher Education* (1991) noted an increasing number of well prepared students that plan to contain educational cost by attending a community college and then transferring to a four year institution.

Further, student expectations of community colleges are rising. Students have become more discriminating consumers who expect more value and services from colleges. An increased emphasis on technology is also placing financial pressures on higher education. Adding to these financial pressures is the fact that higher education processes are often labour intensive. Heterick (1993) estimated that 80 per cent or more of the operations budget is allocated to personnel services.

Re-engineering emerged as a discipline in the 1980's to meet similar challenges in industry. In their book, *Re-engineering the Corporation*, Michael Hammer and James Champy (Hammer and Champy, 1993) proposed that business throw out the old notions of how it should be run, abandon organizational and operational principles and procedures currently in place, and create entirely new ones.

Their formal definition for this new business model was "the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance, such as cost, quality, service, and speed." Others, for example Manganelli & Klein (1994), Carr & Johansson (1995), and Lowenthal (1994), joined the initiative by developing procedures and methods for applying Re-engineering and expanding the concepts based on hard earned experience.

Many Post-secondary institutions have recently turned to Re-engineering as a strategy to meet their new challenges while containing costs (Penrod and Dolence, 1992). A quick survey through college and university web sites as well as recent literature results in a large list of institutions embarking on Re-engineering programmes.

Both Glendale Community College (GCC) and Oklahoma City Community College (OCC) have undertaken Re-engineering projects, termed process engineering programmes, enabling the colleges to serve more students within current space, time and staffing resources allocation by dramatically improving their processes.

GCC's project, the Process Engineering programme (PEP), applies the discipline of systems engineering to processes, thus creating a highly structured methodology; employs a pro-active change management programme; uses existing college resources; and integrates the programme into planning and management activities. OCC's Project VIRGIL utilizes a highly-structured design model, influenced by cross-functional teams. This paper describes the methods and tools of process engineering as applied to these educational institutions and discusses the lessons learned by both organizations. The following discussion outlines the process engineering methodology. An overview is then presented of the each college's programme.

FUNDAMENTAL MEASUREMENT

Intelligence test scores with which arithmetic could be done were called for by Terman and Thorndike before 1920 and required for measurement by

Thurstone in the 1920's. Thurstone (1925) constructed some rather good interval scales with his absolute scaling. With his Law of Comparative Judgement he did even better, producing results that are successful instances of fundamental measurement (Thurstone, 1927). Insistence on order and additivity recurs in Guilford's (1936) definition of measurement. The significant consequence of additivity is the maintenance of the unit of measurement and hence of the invariance of comparisons of measures across the scale. During 1940's Guttman realized that a test score would be ambiguous in meaning unless the total score on the items was sufficient to reproduce the response pattern the score represented.

This led Guttman (1950) to a criterion for judging whether data were good enough to build a scale. The data must demonstrate a joint order shared by items and persons. During 1950's the Danish mathematician George Rasch found that he could not obtain an invariance of test item characteristics over variations in persons unless he could represent the way persons and items interacted to produce a response in a special way.

It is possible to represent a response by an exponential function in which the participation of person and item parameters could have a linear form (Rasch, 1960, p. 120).

Rasch also noted that invariance could be maintained only when data could be gathered so that they cooperated with a stochastic response model that produced a joint order of response probabilities—similar to the joint order Guttman called for (Rasch, 1960, p. 117).

As he worked with this first “Rasch model,” Rasch discovered that the best (*i.e.*, minimally sufficient) statistics from which to estimate person measures and item calibrations were none other than the good old unweighted sums of right answers for persons and for items—the familiar raw scores.

Then Luce and Tukey (1964) showed that an additivity just as good for measurement as that produced by physical concatenation could be obtained from responses produced by the interaction of two kinds of objects (*e.g.*, persons and items), if only this interaction were conducted so that its outcomes (*e.g.*, item response data) were dominated by a linear combination of the two kinds of quantities implied (*e.g.*, the differences between person measures and item calibrations).

Luce and Tukey showed that if care taken; their “conjoint measurement” could produce results as fundamental as Campbell's fundamental measurement. For Luce and Tukey, the moral seems clear: *when no natural concatenation operation exists, one should try to discover a way to measure factors and responses (e.g., gather data) such that the ‘effects’ of different factors are additive* (p. 4).

The realization that Rasch's models could do this when data were collected carefully enough followed (Brogden, 1979; Fischer, 1968; Keats, 1967). Perline, Wright and Wainer (1979) provide empirical demonstrations of the efficacy of the Rasch process in constructing fundamental measurement.

When Andersen (1977) showed that the sufficient statistics that enable fundamental measurement depend on scoring successive categories with the equivalent of successive integers—exactly the ordered category scoring most widely used on intuitive grounds.

The Rasch model for dichotomously scored items was extended to response formats with more than two ordered categories (Andrich, 1978; Wright & Masters, 1982).

Common Practice

As this brief history shows, there has been enough successful theoretical and practical work on the nature and implementation of fundamental measurement to establish its necessity as a basic tool of science and its ready accessibility for educational researchers.

Fundamental measurement is obtained from educational data and, in an intuitive form, educational researchers have relied on it for a long time. In their use of unweighted raw scores and successive integer category weights, educational researchers have been practicing the scoring required for fundamental measurement all along.

Intuitive rather than explicit reliance, however, has meant that they have neither recognized nor enjoyed the benefits of fundamental measurement, and they have not built on its strengths to improve their understanding of education.

Illiterate Theory

The ready accessibility of fundamental measurement and its necessity for the successful practice of science is still unknown to most educational researchers and misunderstood by most psychometricians.

In spite of 60 years of literature explaining the strong reasons for and illustrating the successful application of fundamental measurement in educational research, few psychometricians and fewer educational researchers attempt to construct fundamental measures.

Much of the recent psychometric writing and practice goes on as though no knowledge concerning the theory and practice of fundamental measurement existed. In books claiming to provide the latest ways to make measurements with educational data, one might expect that an exposition of fundamental measurement would not only appear but would dominate the discussion, the choice of methods advanced, and the choice of data analysed. In *Item Response Theory* by Hulin, Drasgow and Parsons and *Applications of Item Response Theory*, edited by Hambleton, there is no discussion of the nature, meaning, or practice of fundamental measurement.

Indeed, a theory of measurement is denied in Hulin, Drasgow and Parsons, who say “in social science we have no well-articulated meta-theory that specifies rules by which one can decide among competing (item response) theories on the basis of the propositions, assumptions, and conclusions of the theories” (p. vii).

The possibility of a relevant theory is despaired of in Hambleton, where the “ineluctable conclusions” are “that no unidimensional item response model is likely to fit educational achievement data” and “the Rasch model is least likely to fit” (Traub, in Hambleton, p. 65).

Response models for one, two and three item parameters are described in detail but their motivation through Thurstone’s (1925) assumption of a latent response process of continuous distribution is unnecessarily abstract and circuitous. It would be so much better education and science to begin with a binomial model for the observable dichotomy as Rasch does (1960, pp. 73). That simple story is not only easier to follow but leads directly to fundamental measurement.

These two books are rife with the usual misunderstandings. The three parameter model is called a logistic model although, when guessing varies, Birnbaum (1968, p. 432) says it is not. The estimation virtues extolled by Swaminathan in a chapter featuring the three parameter model apply, in fact, only to the one parameter Rasch model. Indeed “the likelihood function (of the three parameter model) may possess several maxima” and its value at infinite ability “may be larger than the maximum value found” when ability is finite.

The Rasch model is referred to as a special case of the three parameter model when, in fact, its singular significance for measurement is that it is a unique (necessary and sufficient) deduction from the (fundamental) measurement requirements of joint order and additivity.

The sum of discrimination estimates for items answered correctly, which cannot be a statistic because a statistic must be a function of data and not of other statistics, leads a double life. It is called a sufficient statistic (Bejar, in Hambleton, p. 11; Swaminathan, in Hambleton, p. 30) and then shown to be insufficient because “it is not possible to extend the conditional (estimation) approach to the two parameter model” so that “sufficient statistics exist for ability... only in the Rasch model” (Swaminathan, in Hambleton, p. 36).

Bejar worries that “the logistic model ignores the difficulty of the items answered correctly in assigning a score” (Bejar, in Hambleton, p. 14). But that is exactly what raw scores have always done. Although the score pays close attention to the difficulty of the test, it must be indifferent to items within the test which are answered correctly. This indifference is a necessary consequence of the local independence assumed by all of these models and required for any item banking.

The study of items which a person answers correctly, that is, the investigation of the joint order between observed person responses and estimated item difficulties, is the study of person fit. Does the person’s data fit the measurement model? Is the person’s performance valid? Indifference as items which within a test are answered correctly is necessary for building measurement. But the measure constructed from this indifference is only valid when the person’s performance pattern is stochastically consistent (jointly ordered) with the items’ difficulties.

Impractical Practice

Methods for using these response models are left to computer programmes, mostly LOGIST. Lord help the poor researcher who hasn't got his LOGIST working.

There are no explicit procedures, detailed examples, or even estimation equations provided. Any reader hoping to learn how to apply IRT will have to look elsewhere (*e.g.*, Rasch, 1960; Spearritt, 1982; Wright & Stone, 1979). Even in Hambleton's Chapter, which is dedicated to LOGIST, there are no instructions for how to use it.

The basic problems with the two and three parameter models are made plain by the steps taken to deal with their symptoms. During "estimation in the two and three parameter models... the item parameter estimates drift out of bounds" (Swaminathan, in Hambleton, p. 34). Thus "Range restrictions (must be) applied to all parameters except the item difficulties" to control "the problem of item discriminations going to infinity" (Wingersky, in Hambleton, pp. 47-48). Worse than that, "items with vastly different discrimination and difficulty parameters (can) nonetheless have virtually identical ICCs in the ability interval" where the data are present and the estimation work is to be done (Hulin, Drasgow, & Parsons, p. 100).

All this happens because an empirical binomial ICC expressed on a scale that must be defined by the same data contains only enough information to identify one item parameter. The only way more item parameters can be estimated is to assume that the particular persons participating in the item calibration are random examples of some "true" distribution of ability for which these items are always to be used.

The problem is aggravated by the way these models use ability in two ways at once. The first use is as a "difference" that specifies a distance between person's ability and item difficulty. This difference is essential for the construction of measurement. The second use is as a "factor" to multiply item discrimination so that a different unit can be specified for each item.

This interacts with the first use to confound estimation and prevent the construction of joint order or additivity. The non-linear combination of item parameters prevents the algebraic separation of difficulty and discrimination and hence the derivation of sufficient statistics for estimating them.

When ICC's are allowed to cross (*e.g.*, Hambleton, pp. 45, 46, 163), the manifest difficulty order items varies with ability. This prevents the construction of a variable that can be defined in any general way by the relative difficulties of its items.

The estimations of item discrimination and person's ability are based on a feedback between (i) summing the product of observed response and current ability estimates over persons and (ii) summing the product of observed response and current discrimination estimate over items (Birnbbaum, 1968, pp. 421-422). This process cannot converge because the cumulative effect of the feedback between ability and discrimination pushes their estimates to infinity.

Yen describes the kind of trouble this can get one into: The biggest surprise occurring with the CTBS/U interlevel linking related to the type of scale produced.

As grade increased, the scale scores had decreasing standard deviations and corresponding decreasing standard errors of measurement and increasing item-discriminations.... This result was unexpected because the scaling procedure used with previous tests, Thurstone's absolute scaling, produced a scale with standard deviations increasing with grade. (Yen, in Hambleton, p. 139).

When the response model allows person's ability and item discrimination to interact, this is bound to happen. The decreasing standard deviations are not describing children on an interval scale. They are describing the consequences of an estimation procedure that cannot keep item discriminations from drifting towards infinity.

Parameterizing guessing is much prized in theory. But "attempts to estimate the guessing parameter... are not usually successful" (Hulin, Drasgow, & Parsons, p. 63). In one study, "40per cent of the guessing parameter estimates did not converge even with a sample size of 1593" (Ironson, in Hambleton, p. 160). Even when some estimates of guessing are obtained there are problems.

"If a test is easy for the group (from which guessing parameters are estimated) and then administered to a less able group, the guessing parameters (from the more able group) may not be appropriate" (Wingersky, in Hambleton, p. 48). "When dealing with three parameter logistic ICCs, a non-zero guessing parameter precludes a convenient transformation to linearity" (Hulin, Drasgow, & Parsons, p. 173).

None of this should be least surprising. The formulation of the model shows quite plainly that the explicit interaction between guessing and ability must make any guessing estimates inextricably sample dependent.

Analysis of Fit

The *analysis of fit* for persons and items is addressed at length—two chapters in Hulin, Drasgow, and Parsons, and three in Hambleton. The techniques mentioned and their sources are basically the same. But the substantive discussion and motivation are much better in Hulin, Drasgow, and Parsons. Unfortunately, as with estimation, the methods described are made to seem unnecessarily complex and arcane. If a reader want to apply one of them, he or she will have to go elsewhere to learn how.

Harnisch and Tatsuoaka call for person fit statistics with standard normal distributions and no correlations with ability (in Hambleton, pp. 114, 117-119). That is a good idea when fine tuning fits statistics against data simulated to fit. But they use their ideal to evaluate fit statistics applied to real data. With real data one would prefer fit statistics that are skewed by misfit and correlate negatively with ability (to detect the guessing of low ability persons). But these criteria are the opposite of the ideal Harnisch and Tatsuoaka use in their comparisons.

A special technique for fit analysis used by Yen (in Hambleton, p. 126) on CTB items seems attractive at first, but there is good reason to avoid it. Yen's first step is to smooth out the misfit occurring in her data by regressing observed scores of ability-groups on a monotonic function of their estimated expectations.

Then she makes her "fit" comparisons, not against the regression residuals into which the misfit has just been pushed, but against the regression predictions from which the misfit has just been removed. The reason her "fit statistic is more stable" is that she is no longer analysing misfit.

Ironson does a chapter (in Hambleton, Chapter) on item bias, but the topic is better dealt with in Chapter of Hulin, Drasgow, and Parsons. The trouble with both approaches to bias is that bias found for groups is never uniformly present among members of the group or uniformly absent among those not in the group. For the analysis of item bias to do individuals any good, say, by removing the bias from their measures, it will have to be done on the individual level of the much more useful person fit analyses described in other chapters.

Test Equating

Test equating appears in Chapter of Hulin, Drasgow, and Parsons, only on the way to connecting tests written in more than one language, whereas Hambleton provides a chapter eleven on equating. Unfortunately item banking, the purpose of test equating, is not mentioned.

Neither book describes the actual details of equating sufficiently well to enable the reader to learn how to do it.

There is a refreshing irony in the various accounts (Hulin, Drasgow, & Parsons, pp. 174, 202; Hambleton, pp. 45-46, 132, 178, 181-182; 191) of how equating with the three parameter model is actually accomplished. First of all, whereas the three parameter model is always claimed, the two parameter model is the one actually used to calibrate most of the tests to be equated. This happens because the guessing parameter proves too elusive to be kept as a variable. Then, when it comes to actually connecting two tests, even item discrimination is given up.

The actual equating is based entirely on item difficulty, the one item parameter of the Rasch model. In other words, whenever these authors get involved in actually building a measuring system, even of only two tests, they are forced by what then happens to them to use the only response model that can build fundamental measurement, namely the one item parameter or the Rasch model.

The irony is particularly poignant in the Cook and Eignor attempt to distinguish between three methods for equating tests (in Hambleton, p. 191). Test equating must result in a single linear system of item calibrations and person measures (associated with test scores), which is invariant (but for scale and origin) over the data and also over the methods used to construct this system. Unless all three methods produce the same result, none of them has worked.

These authors claim concern with advancing science by providing better methods for building knowledge. But they offer models for item response data that systematically prevent the construction of fundamental measurement. The only way educational researchers can hope to build knowledge is to use models that insist on data from which fundamental measurement could be constructed—models that contain in their formulation the joint order and additivity conditions required for its construction.

For most of the models presented in these books to produce meaningful item estimates, it is necessary to assume that the persons who provided the data are random events sampled from the “right” simple standard distribution. Persons do not appear as individuals to be measured in these models. Person parameters are not even subscripted in most presentations. The scale reported is routinely standardized to give the sample person a mean of zero and a variance of one as though those particular persons were a true random sample of exactly the population of persons with whom the scale would always be used—and as though any individual person subsequently measured could be usefully understood as no more than a random instance of the particular sample of other persons with whom the item analysis was done.

Figments of Despair

Despair is the latent message in books like this—despair of ever constructing any good mental or psychological variables. Review the testimony. The item response models described are said to be:

- *Hard to Understand:* The “procedures involved in estimation are complex and require sophistication on the part of the user” (Bejar, in Hambleton, p. 3).
- *Difficult to Use:* “The difficulty in applying these models is stressed” (Bejar, in Hambleton, p. 1). “Working with IRT is an arduous process” and “the largest hurdle is the estimation” (Bejar, in Hambleton, p. 3). Would-be users are warned that they “must be ready to pay... by investing substantial resources in parameter estimation and model monitoring” (Bejar, in Hambleton, p. 17).
- *Demanding of large samples of persons and many test items* (Bejar, in Hambleton, p. 3; Ironson, in Hambleton, p. 160; Wingersky, in Hambleton, p. 46). “Even long tests and large samples do not necessarily allow accurate estimation of the guessing parameter” (Hulin, Drasgow, & Parsons, p. 100).
- *Unreliable in Action:* Discrimination eludes capture because of its interaction with ability. Guessing can’t be laid hold of, not only because it is persons and not items who are doing it, but because when plausible estimates do emerge they are sample dependent.
- *Hopeless in any Event:* “Multiple-choice items... are unlikely to be modelled very well by any unidimensional item response model”

(Traub, in Hambleton, p. 57). “Speeded administration and an item format that permits guessing will each introduce another trait into the response process” (Traub, in Hambleton, p. 62). “Data from examinees who vary in their guessing propensities will result in systematically biased calibrations of items and systematically biased estimates of examinee’s ability” (Traub, in Hambleton, p. 63). As for using a model to select data good enough to make measurements with, “It will be a sad day indeed when our conception of measurable educational achievement narrows to the point where it coincides with the criterion of fit to a unidimensional item response model” (Traub, in Hambleton, p. 64).

If this were all we had to look forward to, the future of educational research would be dim indeed. Fortunately there is a readily accessible and remarkably hearty antidote to all the confusion and despair, even to Traub’s terrors.

Foundations for Hope

Hope for the future of measurement in educational research can be found in the following discussion:

- *The theory of fundamental measurement*, the scoring for which we have been engaged in for decades, and the advantages of which we are poised to enjoy. This theory can be put in a form that is easy to understand. The unique response model for applying it, the one parameter Rasch model, is easy to grasp. Finally, what can be more straightforward than seeking the kind of person—item interactions that can be understood as showing us the difference between the person’s ability and the item’s difficulty in an orderly and uniform way?
- *The One-parameter Model is Easy to Use*: No special computer or computer programme is required. Scores of researchers (including R.L. Thorndike) have long since written their own Rasch programmes for their preferred computers and hand calculators. If the PROX approximation is used (Wright & Stone, 1979), the job can be done by hand.
- *Even the Smallest Data Sets can be Usefully Addressed with this Model*: Wright and Stone (1979) get important information about the structure of Knox’s Cube Test and the status of the children tested from no more than 14 items taken by 34 children.
- *The Functioning of this Model is Robust*: The Rasch model is so robust, in fact, that it is routinely used during every second cycle of LOGIST to resist the digressive pressure from the guessing and discrimination parameter estimates to stray (Wingersky, in Hambleton, p. 47). Whenever tests are to be equated, it is the Rasch item parameter, difficulty, that is used (Hulin, Drasgow, & Parsons, pp. 174, 202; in Hambleton, pp. 45-46, 132, 178, 181-182, 191). This not only testifies in a natural way to the necessity of the Rasch model for the construction

of measurement systems, but its many successes in practice (Wright & Bell, 1984) encourage us that there is something useful we can do.

- *The Rasch Sufficient Statistics are Quite Relevant and Hopeful as any Raw Score Ever was*: That's because they come directly from raw scores. If the millions of tests given in the past 70 years were good for anything, if any test ever given and then summarized by the count of right answers was useful, then there is hope.

This hope is particularly strong today because the one parameter model gives us a clear and simple prescription for what a raw score is supposed to do.

This puts us in the position of being able to review every performance pattern before we draw any final conclusions concerning its raw score or implied measure.

Routine analyses of the joint order between estimated item difficulties and observed responses enable us to validate our measures (and calibrations) when they deserve it and to identify, diagnose, and learn from those situations where our analyses point out improbable inconsistencies.

The one parameter model has all the statistical virtues extolled by Swaminathan (in Hambleton, pp. 30, 33, 35). It works well on all kinds of data (Hulin, Drasgow, & Parsons, pp. 57, 95, 96; Hambleton, pp. 200, 221, 226). It does better than the three parameter model two out of three times even in the hands of a researcher who wishes with all his heart it were otherwise and finds it "surprising to observe the one parameter model performing better than the three parameter model... since [as he mistakenly believes] there is no theoretical reason to expect such a result" (in Hambleton, pp. 208-209).

Fortunately the despair intimated in books like these and acted out in choices of methods that are unnecessarily complex and incorrigibly ineffective is not all there is.

There is the theory and practice of fundamental measurement ready and waiting for any educational researcher seriously interested in building educational variables in order to measure the status and change of individuals.

There are many Rasch built item banks doing good work (Wright & Bell, 1984). For a recent overview of Rasch thinking and works there is the Australian Council for Educational Research Golden Jubilee Seminar on *The Improvement of Measurement in Education and Sociology* (Spearritt, 1982). Best of all, the University of Chicago Press (1980) has reprinted Rasch's own book on *Probabilistic Models for Some Intelligence and Attainment Tests*. No scholar of quantitative research method, whether student or professor, need remain unlettered in that great book!

Summary Table of the Main Theoretical Frameworks

The following table summarizes the main *theoretical frameworks* behind almost all the theoretical and research work, and the instructional practices in education (one of them being, of course, the practice of assessment). These different frameworks have given rise to interesting debates among scholars.

Topics	Empiricism	Rationalism	Socioculturalism
Philosophical orientation	Hume: British empiricism	Kant, Descartes: Continental rationalism	Hegel, Marx: cultural dialectic
Metaphorical Orientation	Mechanistic/Operation of a Machine or Computer	Organismic/Growth of a Plant	Contextualist/Examination of a Historical Event
Leading Theorists	B. F. Skinner (behaviorism)/ Herb Simon, John Anderson, Robert Gagné: (cognitivism)	Jean Piaget/Robbie Case	Lev Vygotsky, Luria, Bruner/ Alan Collins, Jim Greeno, Ann Brown, John Bransford
Nature of Mind	Initially blank device that detects patterns in the world and operates on them. Qualitatively identical to lower animals, but quantitatively superior.	Organ that evolved to acquire knowledge by making sense of the world. Uniquely human, qualitatively different from lower animals.	Unique among species for developing language, tools, and education.
Nature of Knowledge (epistemology)	Hierarchically organized associations that present an accurate but incomplete representation of the world. Assumes that the sum of the components of knowledge is the same as the whole. Because knowledge is accurately represented by components, one who demonstrates those components is presumed to know	General and/or specific cognitive and conceptual structures, constructed by the mind and according to rational criteria. Essentially these are the higher-level structures that are constructed to assimilate new info to existing structure and as the structures accommodate more new info. Knowledge is represented by ability to solve new problems.	Distributed across people, communities, and physical environment. Represents culture of community that continues to create it. To know means to be attuned to the constraints and affordances of systems in which activity occurs. Knowledge is represented in the regularities of successful activity.
Nature of Learning (the process by which knowledge is increased or modified)	Forming and strengthening cognitive or S-R associations. Generation of knowledge by (1) exposure to pattern, (2) efficiently recognizing and responding to pattern (3) recognizing patterns in other contexts.	Engaging in active process of making sense of ("rationalizing") the environment. Mind applying existing structure to new experience to rationalize it. You don't really learn the components, only structures needed to deal with those components later.	Increasing ability to participate in a particular community of practice. Initiation into the life of a group, strengthening ability to participate by becoming attuned to constraints and affordances.
Features of Authentic Assessment	Assess knowledge components. Focus on mastery of many components and fluency. Use psychometrics to standardize.	Assess extended performance on new problems. Credit varieties of excellence.	Assess participation in inquiry and social practices of learning. Students should participate in assessment process. Assessments should be integrated into larger environment.

CONTROVERSY

Concerns over how best to apply assessment practices across public school systems have largely focused on questions about the use of high stakes testing and standardised tests, often used to gauge student progress, teacher quality, and school-, district-, or state-wide educational success.

No Child Left Behind

For most researchers and practitioners, the question is not whether tests should be administered at all—there is a general consensus that, when administered in useful ways, tests can offer useful information about student progress and curriculum implementation, as well as offering formative uses for learners. The real issue, then, is whether testing practices as currently implemented can provide these services for educators and students.

In the U.S., the No Child Left Behind Act mandates standardised testing nationwide. These tests align with state curriculum and link teacher, student, district, and state accountability to the results of these tests. Proponents of NCLB argue that it offers a tangible method of gauging educational success, holding teachers and schools accountable for failing scores, and closing the achievement gap across class and ethnicity.

Opponents of standardised testing dispute these claims, arguing that holding educators accountable for test results leads to the practice of “teaching to the test.” Additionally, many argue that the focus on standardised testing encourages teachers to equip students with a narrow set of skills that enhance test performance without actually fostering a deeper understanding of subject matter or key principles within a knowledge domain.

High-stakes Testing

The assessments which have caused the most controversy in the U.S., are the use of high school graduation examinations, which are used to deny diplomas to students who have attended high school for four years, but cannot demonstrate that they have learned the required material when writing exams. Opponents say that no student who has put in four years of seat time should be denied a high school diploma merely for repeatedly failing a test, or even for not knowing the required material.

High-stakes tests have been blamed for causing sickness and test anxiety in students and teachers, and for teachers choosing to narrow the curriculum towards what the teacher believes will be tested. In an exercise designed to make children comfortable about testing, a Spokane, Washington newspaper published a picture of a monster that feeds on fear. The published image is purportedly the response of a student who was asked to draw a picture of what she thought of the state assessment.

Other critics, such as Washington State University’s Don Orlich, question the use of test items far beyond standard cognitive levels for students’ age.

Compared to portfolio assessments, simple multiple-choice tests are much less expensive, less prone to disagreement between scorers, and can be scored quickly enough to be returned before the end of the school year.

Standardised tests (all students take the same test under the same conditions) often use multiple-choice tests for these reasons. Orlich criticizes the use of expensive, holistically graded tests, rather than inexpensive multiple-choice “bubble tests”, to measure the quality of both the system and individuals for very large numbers of students. Other prominent critics of high-stakes testing include Fairtest and Alfie Kohn. The use of IQ tests has been banned in some states for educational decisions, and norm-referenced tests, which rank students from “best” to “worst”, have been criticized for bias against minorities. Most education officials support criterion-referenced tests (each individual student’s score depends solely on whether he answered the questions correctly, regardless of whether his neighbours did better or worse) for making high-stakes decisions.

21st Century Assessment

It has been widely noted that with the emergence of social media and Web 2.0 technologies and mindsets, learning is increasingly collaborative and knowledge increasingly distributed across many members of a learning community. Traditional assessment practices, however, focus in large part on the individual and fail to account for knowledge-building and learning in context. As researchers in the field of assessment consider the cultural shifts that arise from the emergence of a more participatory culture, they will need to find new methods of applying assessments to learners.

Assessment in a Democratic School

Sudbury model of democratic education schools do not perform and do not offer assessments, evaluations, transcripts, or recommendations, asserting that they do not rate people, and that school is not a judge; comparing students to each other, or to some standard that has been set is for them a violation of the student’s right to privacy and to self-determination. Students decide for themselves how to measure their progress as self-starting learners as a process of self-evaluation: real lifelong learning and the proper educational assessment for the 21st century, they adduce.

According to Sudbury schools, this policy does not cause harm to their students as they move on to life outside the school. However, they admit it makes the process more difficult, but that such hardship is part of the students learning to make their own way, set their own standards and meet their own goals. The no-grading and no-rating policy helps to create an atmosphere free of competition among students or battles for adult approval, and encourages a positive cooperative environment amongst the student body.

The final stage of a Sudbury education, should the student choose to take it, is the graduation thesis. Each student writes on the topic of how they have prepared themselves for adulthood and entering the community at large. This

thesis is submitted to the Assembly, who reviews it. The final stage of the thesis process is an oral defence given by the student in which they open the floor for questions, challenges and comments from all Assembly members. At the end, the Assembly votes by secret ballot on whether or not to award a diploma.

Assessing ELL Students

A major concern with the use of educational assessments is the overall validity, accuracy, and fairness when it comes to assessing English language learners (ELL). The majority of assessments within the United States have normative standards based on the English-speaking culture, which does not adequately represent ELL populations. Consequently, it would in many cases be inaccurate and inappropriate to draw conclusions from ELL students' normative scores. Research shows that the majority of schools do not appropriately modify assessments in order to accommodate students from unique cultural backgrounds. This has resulted in the over-referral of ELL students to special education, causing them to be disproportionately represented in special education programmes. Although some may see this inappropriate placement in special education as supportive and helpful, research has shown that inappropriately placed students actually regress in progress.

It is often necessary to utilise the services of a translator in order to administer the assessment in an ELL student's native language; however, there are several issues when translating assessment items. One issue is that translations can frequently suggest a correct or expected response, changing the difficulty of the assessment item. Additionally, the translation of assessment items can sometimes distort the original meaning of the item. Finally, many translators are not qualified or properly trained to work with ELL students in an assessment situation. All of these factors compromise the validity and fairness of assessments, making the results not reliable. Nonverbal assessments have shown to be less discriminatory for ELL students, however, some still present cultural biases within the assessment items.

When considering an ELL student for special education the assessment team should integrate and interpret all of the information collected in order to ensure a non biased conclusion. The decision should be based on multidimensional sources of data including teacher and parent interviews, as well as classroom observations. Decisions should take the students unique cultural, linguistic, and experiential backgrounds into consideration, and should not be strictly based on assessment results.

RESEARCH ON ACHIEVEMENT GROUPING AND TRACKING

Unfortunately, the research base on grouping is extremely dated and does not clearly evaluate the four alternative grouping arrangements described in Table. An analysis of the dates of the most recent comprehensive reviews with opposing conclusions (Kulik and Kulik 1987; Slavin, 1987, 1990) illustrates

just how dated the research is. Not one U.S., study included in Kulik and Kulik's (1987) review of 105 studies nor in Slavin's (1987, 1990) reviews of 43 elementary studies and 29 secondary studies was published after the landmark *Marshall v Georgia* ruling in 1985.

Furthermore, only 5 of the 105 studies reviewed by Kulik and Kulik, and only 4 of the 72 studies reviewed in both of Slavin's reviews were published after 1976, the original passage of the *Education for All Handicapped Children Act*. This legislation was probably more influential than any other event in the history of American education in terms of raising the interest of school personnel in better serving the needs of students with disabilities and of low-performing students.

In fact, only 15 per cent of the studies reviewed were published after the *Hobson v Hansen* ruling in 1969. The preponderance of the research is over 30 years old. The abuses of grouping practices that the courts called "tracking" in *Hobson v Hansen* were probably much more common across America before the Hobson ruling than they are today.

The current question of interest to schools generally differs from the researchers' questions. The researchers have generally attempted to isolate the grouping variable from instruction, keeping instruction the same for all groups and changing only the grouping arrangements. The research question is generally: Does achievement grouping improve learning when all groups are taught using the same materials and methods? Few practitioners exist who would expect achievement grouping to have any consistent effect without matching instruction to needs.

The questions of interest to schools include the following:

- Is achievement grouping with appropriately varied instruction for each group more effective than mixed-age, mixed-ability grouping?
- Is achievement grouping with appropriately varied instruction more effective than traditional age-based grouping?

Research that does not attempt to vary instruction appropriately for different grouping arrangements does not answer practitioners' questions about grouping. (See Allan, 1991 and Kulik, 1991 for further details regarding the mismatch between practitioners' and researchers' questions on grouping.) Most of the studies on grouping do not describe at all the nature of the instruction that occurred in the study.

The studies of elementary school grouping alternatives have more complete descriptions of the instruction than the secondary studies of grouping. After using a "best evidence synthesis" to seek out patterns of positive and negative effects in 43 studies comparing elementary school grouping arrangements, Slavin was able to conclude:

"Taken together, the evidence points to a conclusion that for ability grouping to be effective at the elementary level, it must create true homogeneity on the specific skill being taught and instruction must be closely tailored to students' level of performance. (p. 323).

This is consistent with the *Marshall v Georgia* ruling. The courts saw positive effects for ability grouping when the grouping was based on achievement in the specific skills taught in the programme.

Furthermore, Slavin found that the conditions leading to favourable effects for grouping were more common in “within-class” grouping and rarely existed in “between-class” grouping. Within-class grouping involves assigning children to groups within a class. Between-class grouping involves assigning children to classes for the entire year based on their ability or achievement levels. Slavin reasoned that a student’s placement, though optimal for one subject, may not be optimal for another in between-class grouping at the elementary level.

One model, the Joplin Plan, could not be categorized as within-class or between-class grouping. In the Joplin plan, students are grouped into mixed-age mixed-ability classes, then placed in subject-specific achievement groups formed across classes for instruction in reading and/or mathematics. For example, at a common mathematics period, all students might move to a class composed of students at the same performance level in mathematics drawn from different classes and grade levels. One mathematics group might have high first, average second, and low third graders in it, but all would be at the same approximate point in the learning sequence. These instructional groups are also flexible and not permanent. Groupings are frequently reassessed and changed if student performance warrants it. Slavin found a strong positive effect for the Joplin plan.

Based on these findings, Slavin (1991) concludes that for the elementary level he is not opposed to assigning students to mixed-ability classes and grouping children within or across classes into achievement groups when appropriate. He opposes between-class grouping where students are assigned to self-contained classes based on their ability or performance level. At the elementary level, between-class grouping approximates tracking, when the same groups are maintained for instruction in all subjects.

Slavin’s (1990) review of secondary school research was more problematic. He tried again to separate the studies of within-class grouping from those of between-class grouping to determine if the same pattern of results found at the elementary level was also evident at the secondary level. He found no effects for grouping of any kind. It is not surprising that there were no effects for within-class grouping at the secondary level, though there were at the elementary level. Even if secondary teachers divided their classes into smaller groups for instruction, thereby fitting the criteria for the “within-class” grouping arrangement, it is unlikely that they would modify the instruction for each of the small groups, doubling or tripling the number of preps they would have in a day. Each group would receive only 1/3 of the instructional time they would otherwise receive.

That Slavin also found no effect for between-class (assigning students to different classes according to their achievement level) grouping at the secondary level is more surprising. Between-class grouping at the secondary level is as

subject-specific as within-class grouping at the elementary level. Classes are organized by subject at the secondary level, so between-class grouping does not result in students being assigned to the same class for all subjects as it does at the elementary level.

Slavin concludes: “If the effects of ability grouping on student achievement are zero, then there is little reason to maintain the practice... Arguments in favour of ability grouping depend on assumptions about the effectiveness of grouping, at least for high achievers. In the absence of any evidence of effectiveness, these arguments cannot be sustained” (p. 492, 1990).

Slavin’s (1991) suggestion that using cooperative learning with mixed-age mixed-ability groups is more viable than between-class grouping is having profound impact in the restructuring movement. (See *Educational Leadership*’s issue featuring restructuring, March, 1991.) Slavin’s research is frequently cited to support the extensive restructuring of secondary schools to incorporate project-based learning where small mixed-ability cooperative learning groups spend much of their school time working cooperatively on large-scale projects, such as setting up a museum featuring the local community.

However, Slavin’s conclusions regarding between-class achievement grouping at the secondary level are seriously limited by the selection rules he used in his meta-analysis. Slavin systematically eliminated any study that involved different programmes for different levels. Slavin included only experimental studies that compared students at the same grade level taking the same course in achievement-grouped versus non-achievement-grouped classes. For example, only ninth-grade students in Math 9 were compared. Ninth graders taking Algebra or Math 8 would not be compared with ninth-grade students taking Math 9. One treatment would involve high, average, and low sections of Math 9. The other treatment involved all levels mixed in Math 9 classes. Slavin comments regarding this limitation:

“The experimental studies do not compare students in Algebra 1 to those in Math 9, or students who take 4 years of math to those who take 2. The conclusions drawn in this section are limited, therefore, to the effects of between-class grouping *within the same courses*, and should not be read as indicating a lack of differential effects of tracking [or achievement grouping]. (Slavin, 1990, p. 486-7)

This is a major caveat. Most of the practical impact of achievement grouping would be expected to come from high level students taking courses that cover more advanced content. Any studies that would detect this effect were excluded from Slavin’s reviews.

Kulik and Kulik (1991) used different selection criteria for their meta-analyses and ended up including a different set of studies. Very few studies reviewed by Slavin were also reviewed by the Kuliks. In discussing the results of the Kulik and Kulik review (1991), Kulik (1991) distinguished three types of programmes:

Type I: Simple programmes in which all ability groups are taught with the same or similar materials and by the same or similar methods.

Type II: Programmes in which teaching materials and methods are adjusted to meet the special needs of a specific aptitude group (for example, enriched instruction for the talented and gifted).

Type III: Programmes in which adjustment of teaching materials is so extensive that it affects a student's rate of progress through school (for example, programmes of accelerated instruction).

Effects varied according to type, with negligible effects found for Type I programmes (.1 effect size), stronger effects for Type II programmes (.4 effect size), and much stronger effects for Type III programmes (1.0 effect size). Kulik's (1991) conclusions seem to support the practice of achievement grouping as defined by the courts. The more instruction is varied to meet the specific needs of students in the achievement groups, the more effective it is.

However, most of the Type II and Type III research evaluated only programmes for the gifted and high-performing students. As Slavin (1991) points out, evaluating the effects of gifted programmes only on gifted students leaves open the possibility that gifted programmes might have positive effects for all students. Indeed many reformers (*e.g.*, Oakes; see interview with Oakes in O'Neil, 1992) argue that gifted programmes should be offered to all students. However, the effectiveness of gifted programmes for all students was not evaluated in this research. Other research (described later) raises considerable doubt that gifted programmes would have positive effects for all students.

Summary: Flawed research methodology seems to support the conclusion that there is no clear answer to the question: Does achievement grouping improve learning when all groups are taught using the same materials and methods? This is a question few ask. The contradictions in the findings within each meta-analysis seem to indicate that grouping arrangements alone are not the primary variable for school effectiveness. Whether effective practices are used for all levels, particularly the low achievement levels, is the legal test for racial equity. If the learning of low-achieving minority children is accelerated, equity is served. If not, inequity is present.

Research on Mixed-Age Grouping

Pavan (1977) reviewed 51 comparisons of mixed-age grouping conducted between 1968 and 1978 and concluded that *mixed-age* grouping was more effective than *age-based* grouping. Pavan's conclusion was used to support the non-graded model promoted by the National Association for the Education of Young Children (Bredekamp, 1987), which not only mixes ages, but also mixes abilities.

However, Pavan's research does not support mixed-ability grouping within the mixed-age model. The mixed-age models she evaluated included both achievement grouping, as in the Joplin plan, and mixed-ability grouping. Pavan did not break down the results for mixed-age models according to whether achievement grouping or mixed-ability grouping was used. Rather she grouped the effects together.

Gutierrez and Slavin (1992) reviewed Pavan's same data set and more (57 studies), but categorized the studies according to instructional and grouping practices used among the mixed-age models. Their findings did not contradict Pavan's; they also found more positive than negative significant results favouring the mixed-age ("non-graded") model.

However, they found that the models that contributed most to the overall positive effect Pavan found for mixed-age primaries actually used achievement grouping for instruction in reading and/or mathematics (the Joplin plan), not mixed-age mixed-ability grouping, as is promoted by Pavan (1992) and Bredekamp (1987). Gutierrez and Slavin concluded that the "non-graded organization can have a positive effect on student achievement if cross-age grouping is used to allow teachers to provide more direct instruction to students but not if it is used as a framework for individualized instruction" (p. 333).

Achievement grouping across ages, rather than only within grade levels, allows teachers to reduce the number of within-class reading and math groups they teach at any given time, thereby reducing the need for independent seatwork and follow-up. Gutierrez and Slavin (1992) indicated that several evaluators of Joplin-like programmes noted specifically that mixed-age groupings made within-class groupings unnecessary, so teachers could use the entire class period to teach the whole class. Mixed-ability models involved individualized instruction, learning stations, learning activity packets, and other individualized or small group activities which reduced direct instruction time with little corresponding increase in appropriateness of instruction to meet individual needs, according to Gutierrez and Slavin (1991). They point out that the research on non-gradedness has not evaluated the currently popular model promoted by the NAEYC and Katz *et al.* (1991):

The movement towards developmentally appropriate early childhood education and its association with nongrading means that the non-graded primary schools of the 1990s will often incorporate 4- and 5-year-olds (earlier forms rarely did so) and that instruction in non-graded primary programmes will probably be more integrated and thematic, and less academically structured or hierarchical, than other schools.... Whether these models will have positive or negative effects on ultimate achievement is currently unknown. (p. 370)

Anderson and Pavan (1993) later expanded Pavan's original review (1977) of non-graded, or mixed-age primaries, to include 64 studies. They found positive effects for the non-graded model, but again they did not break down the results according to whether the models used mixed-ability or achievement grouping within the mixed-age model. Without this breakdown, their conclusions cannot be used to support mixed-ability grouping practices within the mixed-age model.

Gutierrez and Slavin (1992) also point out an additional problem with the research on non-graded models: If the non-graded model is used to allow students more time to complete the primary grades, as they usually are, then the average "third-year" student may be older in the non-graded school than in the graded school, creating an artificial advantage for the non-graded model in this research literature.

McGurk and Pimentle (1992) also found empirical support for the Joplin plan in their review of the research on mixed-age (non-graded) models. Mixed-age models that did not use the Joplin plan obtained academic achievement that was comparable to the age-based grouping. Pratt (1986) found no consistent advantage for one grouping plan over another in academic achievement, nor did Cotton (1993), Miller (1990; 1991), and Ford (1977). In their review of reviews, Ellis and Fouts (1994) conclude that most reviews find the non-graded primary has no positive effects on achievement.

Summary: The research on mixed-age models includes mixed-ability and achievement grouping within a mixed-age environment. The findings cannot be used to understand the effects of achievement or mixed-ability grouping without separate analysis. Separate analyses indicate that better results are associated with the Joplin plan for achievement grouping. An important question left unanswered in all of these reviews is how well the low-performing students did. As the courts have already ruled, the question is not whether a school groups by ability or not; the question is how well the low-performers do, especially when they include a larger proportion of legally protected minority students. If these low achieving students are not learning as well as they could, equity is not being served, regardless of the grouping arrangement.

Excellence Issues

Our national reform goal is to achieve world class standards. A key recommendation of many organizations leading our national reform efforts is to achieve equity by mixing students with widely differing abilities in the classroom. Achieving world class standards though requires much more. Another approach to resolving the problem of equity is to look for school models where low achievers reach remarkably high performance levels and find reliable ways to replicate those models.

One of the few organizations that has taken a serious look at identifying the best performance in the world is the American Federation of Teachers (AFT). A recent comparison of the achievement levels of lower track students in European countries with American students reveals that lower track students in Europe achieve remarkably high performance levels compared to mainstream students in America (AFT, 1995).

The gateway exams for school completion for lower track students in Europe are much more rigorous than America's comparable exam for a Graduation Equivalency Diploma, which is normed to reflect what 75 per cent of America's high school graduates know by the end of grade 12. At grade 9 or grade 10, 60 per cent to 85.5 per cent of the students in European countries pass their much more rigorous exams. The achievement levels of lower track students in European countries using tracking systems are much higher than the expectations for American students.

Certainly the relatively homogeneous societies of Europe do not face the same equity issues that the racially heterogeneous American society faces. If transferred to America, the more rigid tracking of students into different schools

at an early age and the permanent assignment of students to classroom groups over several years could easily translate into permanently lower expectations for minority children.

Tracking per se is not necessarily the cause of the high performance levels for lower track students in Europe. The American Federation of Teachers suggests other factors leading to the effectiveness of the European system: national or state-administered assessments, strong incentives to excel, and a common curriculum. These aspects of the European model seem crucial if world class excellence is to be achieved.

Can Mixed-Ability Grouping Lead to World Class Achievement?

If mixed-age mixed-ability grouping can result in low achievers reaching the same high performance levels found in Europe, then achievement grouping is not necessary. The fact that this challenge has not been met using mixed-age mixed-ability grouping does not mean that the challenge is impossible to meet. However, there are several requirements that mixed-age mixed-ability grouping must meet in order to make the case that world class excellence can be achieved using mixed-ability grouping.

Does Quality Instruction Look the Same for High and Low-achieving Students?: Mixed-ability grouping assumes that the same kind of instruction is best for achieving excellence with both high and low achievers. In her frequently cited book, *Keeping Track*, Oakes (1985) analysed descriptive data collected on 25 secondary schools during the early 1970's and documented that inferior instruction was still occurring in many schools, in spite of the 1967 and 1969 *Hobson v Hansen* rulings. She judged the instruction for the low groups inferior not because fewer resources were available to these groups, as the courts did. She judged the instruction in the low groups inferior because the quality of instruction was different. Low groups did lots of worksheets, worked alone more, and spent more time reading out of textbooks. The high groups received more experience-based learning and challenging problems that are likely to have more than one right answer (O'Neil, 1992).

Oakes argues that with mixed-ability grouping, all students will have equal access to the higher quality instruction. Her argument assumes that what she has identified as "quality" instruction will have the same beneficial results for both high and low-achievers. Only under this condition is equity achieved by providing the same instruction for all students.

A very recent study by Gamoran, Nystrand, Berends, and LePore (1995) evaluated the effects of various instructional variables on the learning of high and low performing students. They examined the characteristics of students placed in 92 honours, regular, and remedial English classes in eighth and ninth grade, looking at the effects of similarities and differences in the instruction across achievement groups on the learning of these groups. They found that some instructional variables—discussion and authentic questions—had reversed effects on the achievement of different achievement groups:

“This difference [in the levels of discussion across groups] turned out to be potent for achievement inequality, however, because discussion only benefited students in the high-level classes. Authenticity was also consequential for achievement gaps, but not in the way originally expected: It occurred with similar frequency across classes, but it was beneficial to high-ability students and detrimental to those in low-ability classes.” (p. 708)

The finding for discussion “contradicted our expectation that discussion would benefit low-ability students most of all” (p. 706). The finding for authenticity was “not consistent... with our speculation, based on prior research, that authentic discourse offers greater benefits in low-ability classes than elsewhere. We found just the opposite” (p. 706).

Gamoran *et al.*’s study (1995) is important because it raises a crucial question: Does quality instruction look the same for high and low-ability students? If features of quality vary according to the achievement level of the group, then Oakes (1985), and similarly Goodlad’s (1984), argument is flawed. What these researchers thought was a feature of high-quality instruction (authentic questions, open-ended discussion) may actually not represent high quality instruction for students at lower achievement levels. Mixing low achievers with high achievers and providing instruction that benefits only high achievers could have the opposite effect and not increase equity.

Can Nonstandardized Expectations Result in World Class Achievement?: Expectations play an important role in achievement (Means, Moore, Gagne, & Hauck, 1979; Rist, 1970). Different grouping arrangements have strong implications for student expectations. In three of the four models in Table, age-based grouping, tracking, and achievement grouping, expectations can be clearly defined, or standardized, for each group. In mixed-age, mixed-ability grouping, common expectations do not exist for the group, but vary by individual.

When students are grouped by age, all children of the same age face the same grade-level standards and are expected to learn the curriculum provided for that grade level. Early proponents of tracking criticized the appropriateness of age-based expectations (Turney, 1931), just as current advocates of mixed-age, mixed-ability grouping do (Bredenkamp, 1987). Not all children of the same age should be expected to achieve the same outcomes. Tracking redefines expectations for a child’s performance based on the child’s general ability rather than age. Expectations though are still standardized for the different tracks (*e.g.*, European systems).

Achievement grouping temporarily redefines short term expectations based on the current achievement level of the child in the specific subject. All children in a given achievement group generally start from the same place, with different achievement groups starting from different places. Long-term expectations though are generally referenced to the age-level expectations. All achievement groups within the same larger class group work towards achieving, at a minimum, the same long-term expectations defined for that group. Some achievement groups may exceed these standardized expectations.

In mixed-age, mixed-ability grouping expectations vary by individual. The teacher is the judge of what should be expected of each individual and the children are not pressured to achieve expectations that are inappropriate for them (Bredenkamp, 1987). In theory varied expectations for each individual sounds fair and equitable. In reality though, does it work out that way? How does mixed-ability grouping with variable expectations interact with the noted tendency that teachers tend to communicate more positively with children they perceive as bright and more negatively with children they perceive as slow (Cooper, 1979).

Some ethnographic research evaluated the fairness of teachers in varying expectations appropriately in “progressive” schools that emphasized the importance of variable expectations according to the unique abilities of each child (Atkinson, 1985; Bernstein, 1974; Sharp, Green, & Lewis, 1975; Simon, 1981; Willis, 1977). Atkinson (1985) concluded that the shift from traditional to progressive methods in England represented a shift from visible to invisible control.

Sharp, Green, and Lewis (1975) describe how this shift occurs in case studies of three teachers in a model progressive school:

“Whereas all three teachers would claim to be supporters of the egalitarian principle that all pupils are of equal worth, having an equal right to receive an education appropriate to their needs, in practice there was a marked degree of differentiation among the pupils in terms of the amounts and kinds of interaction they had with their teachers....Those pupils whom their teachers regarded as more successful tended to be given far greater attention than the others. The teachers interacted with them more frequently, payed [sic] closer attention to their activities, subtly structuring and directing their efforts in ways which were noticeably different from the relationship with other pupils less favourably categorized.” (p. 115)

The children who received less attention were the lower performing children who were from lower working class families, while the children the teacher spent more time with were higher performing children who were also from a higher social class. These inequities occurred in classrooms using mixed-ability grouping taught by teachers espousing strong beliefs in the egalitarian principles undergirding progressivism.

For example, Michael’s teacher described him as a “peculiar” boy who wants to “go his own sweet way.” The teacher said she would not “force” or “make” Michael do activities, even where his achievement was poor compared with other children, because to do so would violate the integrity of the child. Yet she did say: “But he’s ever so willing to join in if you organize a little group-but he doesn’t *need* to...,” so Michael often was not invited to participate (pp. 137-8, Sharp, Green, & Lewis, 1975).

Similar observations were made by other ethnographic researchers, who also shared the egalitarian goals of progressivism (Atkinson, 1985; Bernstein, 1974; Simon, 1981; Willis, 1977). For example, Willis (1977) concluded: “...it can be

argued that often “progressivism” has had the contradictory and unintended effect of helping to strengthen processes within the counter-school culture which are responsible for the particular subjective preparation of labour power and acceptance of a working class future in a way which is the very opposite of progressive intentions in education.” (p. 178)

Apparently, holding different expectations for different students in the same instructional groups, as is recommended in mixed-age mixed-ability grouping arrangements, can result in a much more insidious form of inequality. When the same expectations are held for all members of the group, as occurs in achievement grouping or age-based grouping arrangements, and even in tracking, the differential expectations for the different groups are at least public and can be agreed upon in a partnership of teachers, parents, and children. The openness of the expectations for each group is possibly more democratic than the veiled nature of a teacher’s arbitrary, personal expectations for each student in a mixed-age mixed-ability group. At least, one certainly cannot simply assume that equity will be better served by mixed-age, mixed-ability grouping.

An important point that seems often overlooked is that a model that emphasizes variable expectations for each individual student is also incompatible with our national goal to establish standards. In reconciling the NAEYC’s non-graded, mixed-ability model, which emphasizes developmentally appropriate expectations, with the national movement to establish standards, the NAEYC advocates that governing bodies redefine standards to mean not what students should be able to do, but how teachers should teach.

Does Mixed-ability Grouping Raise Self-esteem?: If it does, the next question is whether higher self-esteem significantly contributes to excellence. A major criticism of achievement grouping is that it lowers the self-esteem of students in low-achievement groups. Kulik and Kulik (1982) and Kulik (1985) reviewed the research regarding effects of grouping on attitude and self-esteem. They found that achievement grouping in a subject resulted in a better attitude towards that subject but did not change attitudes about school.

In regard to self-esteem, the Kuliks’ findings contradict the prevailing expectation. Achievement grouping into high, average, and low groups had a small overall effect on self-esteem, but effects tended to be slightly positive for low-achievement groups and slightly negative for high and average ones (Kulik & Kulik, 1982; Kulik, 1985). Limited studies of remedial programmes indicate that achievement grouping has positive effects on the self-esteem of slow learners (Kulik, 1985). Vaughn (in press) has found similar results in a longitudinal study. Self-esteem decreased for children who moved from the low achievement group into mixed-ability classes.

Allan (1991) asked Kulik for a possible explanation for this surprising result: “Kulik (personal communication) raises an interesting point on the relative importance of the effects of labelling versus the effects of daily classroom experience. He suggests that the labelling (by placement of a student into a low-medium-high group) may have some transitory impact on self-esteem but

that impact may be quickly overshadowed by the effect of the comparison that the student makes between himself or herself and others each day in the classroom. Low-ability students may experience feelings of success and competency when in a classroom with others of like ability, and high-ability students may encounter greater competition for the first time. While the data cannot, in themselves, identify the cause of these findings, the results make it clear that we must re-examine the arguments about self-esteem in light of them.” Other research is often cited to contradict these conclusions. Analyses of the effects of the non-graded primary on self-esteem and attitude frequently find that the non-graded primary has positive effects on both (Ford, 1977; Johnson, Johnson, Pierson, & Lyons, 1985; Miller, 1990; Pavan, 1977; Pratt, 1986; Way, 1981). However, as noted earlier, the non-graded model has included both mixed-age achievement grouping, as in the Joplin plan, and mixed-age mixed-ability grouping. The findings do not necessarily indicate that the models that mixed abilities caused these effects.

In the evaluation of Project Follow Through, the largest educational study ever funded by the U.S., Department of Education, Associates reported very surprising results for self-esteem (1977). The most effective model, which used achievement grouping, produced the largest effects for self-esteem, indicating that self-esteem may be more a function of successful learning than grouping arrangement.

“The performance of Follow Through children in the Direct Instruction sites on the affective measures is an unexpected result. The Direct Instruction Model does not explicitly emphasize affective outcomes of instruction, but the sponsor has asserted that they will be consequences of effective teaching. Critics of the model have predicted that the emphasis on tightly controlled instruction might discourage children from freely expressing themselves, and thus inhibit the development of self-esteem and other affective skills. In fact, this is not the case.” (Abt, IV-B, 1977, p. 73)

The five major models evaluated in Project Follow Through claiming self-esteem as an important goal actually resulted in more negative effects for self-esteem when compared to traditional models of schooling.

How do we know when Equity has been Served?

To argue that separating children by achievement levels denies them equity in education assumes that the classroom is much like a bus: If students have equal access to a seat in the classroom, equity has been served. Equity in education requires more. Equity is clearly served when the achievement of minority children matches the best achievement in the world.

Equity is clearly served when the growth rates of children starting at low achievement levels matches or exceeds the growth rates of children starting at high achievement levels. By observing closely when these events occur, educators may learn more about what it takes to achieve excellence with equity. The critical variables have more to do with instruction than with grouping.

4

Test Assessment

A test or examination is an assessment intended to measure a test-taker's knowledge, skill, aptitude, physical fitness, or classification in many other topics (*e.g.*, beliefs). A test may be administered orally, on paper, on a computer, or in a confined area that requires a test taker to physically perform a set of skills. Tests vary in style, rigour and requirements. For example, in a closed book test, a test taker is often required to rely upon memory to respond to specific items whereas in an open book test, a test taker may use one or more supplementary tools such as a reference book or calculator when responding to an item.

A test may be administered formally or informally. An example of an informal test would be a reading test administered by a parent to a child. An example of a formal test would be a final examination administered by a teacher in a classroom or an I.Q. test administered by a psychologist in a clinic. Formal testing often results in a grade or a test score. A test score may be interpreted with regards to a norm or criterion, or occasionally both. The norm may be established independently, or by statistical analysis of a large number of participants. A standardised test is any test that is administered and scored in a consistent manner to ensure legal defensibility. Standardised tests are often used in education, professional certification, psychology (*e.g.*, MMPI), the military, and many other fields.

A non-standardised test is usually flexible in scope and format, variable in difficulty and significance. Since these tests are usually developed by individual instructors, the format and difficulty of these tests may not be widely adopted or used by other instructors or institutions. A non-standardised test may be used to determine the proficiency level of students, to motivate students to study,

and to provide feedback to students. In some instances, a teacher may develop non-standardised tests that resemble standardised tests in scope, format, and difficulty for the purpose of preparing their students for an upcoming standardised test.

Finally, the frequency and setting by which a non-standardised tests are administered are highly variable and are usually constrained by the duration of the class period. A class instructor may for example, administer a test on a weekly basis or just twice a semester. Depending on the policy of the instructor or institution, the duration of each test itself may last for only five minutes to an entire class period. In contrasts to non-standardised tests, standardised tests are widely used, fixed in terms of scope, difficulty and format, and are usually significant in consequences. Standardised tests are usually held on fixed dates as determined by the test developer, educational institution, or governing body, which may or may not be administered by the instructor, held within the classroom, or constrained by the classroom period. Although there is little variability between different copies of the same type of standardised test (*e.g.*, SAT or GRE), there is variability between different types of standardised tests.

Any test with important consequences for the individual test taker is referred to as a high-stakes test. A test may be developed and administered by an instructor, a clinician, a governing body, or a test provider. In some instances, the developer of the test may not be directly responsible for its administration. For example, Educational Testing Service (ETS), a nonprofit educational testing and assessment organisation, develops standardised tests such as the SAT but may not directly be involved in the administration or proctoring of these tests. As with the development and administration of educational tests, the format and level of difficulty of the tests themselves are highly variable and there is no general consensus or invariable standard for test formats and difficulty. Often, the format and difficulty of the test is dependent upon the educational philosophy of the instructor, subject matter, class size, policy of the educational institution, and requirements of accreditation or governing bodies. In general, tests developed and administered by individual instructors are non-standardised whereas tests developed by testing organisations are standardised.

History

Ancient China was the first country in the world that implemented a nationwide standardised test, which was called the imperial examination. The main purpose of this examination was to select for able candidates for specific governmental positions. The imperial examination was established by the Sui Dynasty in 605 AD and was later abolished by the Qing Dynasty 1300 years later in 1905.

England had adopted this examination system in 1806 to select specific candidates for positions in Her Majesty's Civil Service. This examination system was later applied to education and it started to influence other parts of the world as it became a prominent standard (*e.g.*, regulations to prevent the markers from knowing the identity of candidates), of delivering standardised tests.

Influence of World Wars on Testing

Both World War I and World War II made many people realise the necessity of standardised testing and the benefits associated with these tests. One main reason people saw the benefits was from the Army Alpha and Army Beta tests, which were used during WWI to determine human abilities. Alongside the Army Alpha, the Stanford-Binet Intelligence Scale “added momentum to the testing movement.” Soon after, colleges and industry began using tests to help in accepting and hiring people based on performance of the test. Another reason more tests began to come forth was that people were realising that the distance between secondary education and higher education was widening after WWII. In 1952, the first Advanced Placement (AP) test was administered to begin closing the gap between high schools and colleges.

Modern Day use of Tests

Education

Some countries such as the United Kingdom and France require all their secondary school students to take a standardised test on individual subjects such as the General Certificate of Secondary Education (GCSE) (in England) and Baccalauréat respectively as a requirement for graduation. These tests are used primarily to assess a student’s proficiency in specific subjects such as mathematics, science, or literature. In contrast, high school students in other countries such as the United States may not be required to take a standardised test to graduate.

Moreover, students in these countries usually take standardised tests only to apply for a position in a university programme and are typically given the option of taking different standardised tests such as the ACT or SAT, which are used primarily to measure a student’s reasoning skill. High school students in the United States may also take Advanced Placement tests on specific subjects to fulfil university-level credit. Depending on the policies of the test maker or country, administration of standardised tests may be done in a large hall, classroom, or testing centre. A proctor or invigilator may also be present during the testing period to provide instructions, to answer questions, or to prevent cheating.

Grades or test scores from standardised test may also be used by universities to determine if a student applicant should be admitted into one of its academic or professional programmes. For example, universities in the United Kingdom admit applicants into their undergraduate programmes based primarily or solely on an applicant’s grades on pre-university qualifications such as the GCE A-levels or Cambridge Pre-U. In contrast, universities in the United States use an applicant’s test score on the SAT or ACT as just one of their many admission criteria to determine if an applicant should be admitted into one of its undergraduate programmes.

The other criteria in this case may include the applicant's grades from high school, extracurricular activities, personal statement, and letters of recommendations. Once admitted, undergraduate students in the United Kingdom or United States may be required by their respective programmes to take a comprehensive examination as a requirement for passing their courses or for graduating from their respective programmes.

Standardised tests are sometimes used by certain countries to manage the quality of their educational institutions. For example, the No Child Left Behind Act in the United States requires individual states to develop assessments for students in certain grades. In practice, these assessments typically appear in the form of standardised tests. Test scores of students in specific grades of an educational institution are then used to determine the status of that educational institution, *i.e.*, whether it should be allowed to continue to operate in the same way or to receive funding.

Finally, standardised tests are sometimes used to compare proficiencies of students from different institutions or countries. For example, the Organisation for Economic Co-operation and Development (OECD) uses Programme for International Student Assessment (PISA) to evaluate certain skills and knowledge of students from different participating countries.

Licensing and Certification

Standardised tests are sometimes used by certain governing bodies to determine if a test taker is allowed to practice a profession, to use a specific job title, or to claim competency in a specific set of skills. For example, a test taker who intends to become a lawyer is usually required by a governing body such as a governmental bar licensing agency to pass a bar exam.

Immigration and Naturalisation

Standardised tests are also used in certain countries to regulate immigration. For example, intended immigrants to Australia are legally required to pass a citizenship test as part of that country's naturalisation process.

Competitions

Tests are sometimes used as a tool to select for participants that have potential to succeed in a competition such as a sporting event. For example, serious skaters who wish to participate in figure skating competitions in the United States must pass official U.S., Figure Skating tests just to qualify.

Group Memberships

Tests are sometimes used by a group to select for certain types of individuals to join the group. For example, Mensa International is a high I.Q. society that requires individuals to score at the 98th percentile or higher on a standardised, supervised IQ test.

Types of Tests

Written Tests

Written tests are tests that are administered on paper or on a computer. A test taker who takes a written test could respond to specific items by writing or typing within a given space of the test or on a separate form or document. In some tests; where knowledge of many constants or technical terms is required to effectively answer questions, like Chemistry or Biology - the test developer may allow every test taker to bring with them a cheat sheet. A test developer's choice of which style or format to use when developing a written test is usually arbitrary given that there is no single invariant standard for testing.

Be that as it may, certain test styles and format have become more widely used than others. Below is a list of those formats of test items that are widely used by educators and test developers to construct paper or computer-based tests. As a result, these tests may consist of only one type of test item format (*e.g.*, multiple choice test, essay test) or may have a combination of different test item formats (*e.g.*, a test that has multiple choice and essay items).

Multiple Choice

Multiple choice is a form of assessment in which respondents are asked to select the best possible answer (or answers) out of the choices from a list. The multiple choice format is most frequently used in educational testing, in market research, and in elections, when a person chooses between multiple candidates, parties, or policies. Multiple choice testing is particularly popular in the United States.

Although E. L. Thorndike developed an early multiple choice test, Frederick J. Kelly was the first to use such items as part of a large scale assessment. While Director of the Training School at Kansas State Normal School (now Emporia State University) in 1915, he developed and administered the Kansas Silent Reading Test. Soon after, Kelly became the third Dean of the College of Education at the University of Kansas. The first all multiple choice, large scale assessment was the Army Alpha, used to assess the intelligence of World War I military recruits.

The items of a multiple choice test are often colloquially referred to as "questions," but this is a misnomer because many items are not phrased as questions. For example, they can be presented as incomplete statements, analogies, or mathematical equations. Thus, the more general term "item" is a more appropriate label. Items are stored in an item bank.

Structure

Multiple choice items consist of a stem and a set of options. The *stem* is the beginning part of the item that presents the item as a problem to be solved, a question asked of the respondent, or an incomplete statement to be completed,

as well as any other relevant information. The options are the possible answers that the examiner can choose from, with the correct answer called the *key* and the incorrect answers called *distractors*. Only one answer can be keyed as correct.

This contrasts with multiple response items in which more than one answer may be keyed as correct. Usually, a correct answer earns a set number of points towards the total mark, and an incorrect answer earns nothing. However, tests may also award partial credit for unanswered questions or penalise students for incorrect answers, to discourage guessing. For example, the SAT removes a quarter point from the test taker's score for an incorrect answer. For advanced items, such as an applied knowledge item, the stem can consist of multiple parts. The stem can include extended or ancillary material such as a vignette, a case study, a graph, a table, or a detailed description which has multiple elements to it. Anything may be included as long as it is necessary to ensure the utmost validity and authenticity to the item. The stem ends with a lead-in question explaining how the respondent must answer. In a medical multiple choice items, a lead-in question may ask "What is the most likely diagnosis?" or "What pathogen is the most likely cause?" in reference to a case study that was previously presented.

Examples

If, $a = 1$, $b = 2$. What is $a+b$?

- A. 12
- B. 3
- C. 4
- D. 10
- E. 8

In the equation $2x + 3 = 4$, solve for x .

- A. 4
- B. 10
- C. 0.5
- D. 1.5
- E. 8

Ideally, the MCQ should be asked as a "stem", with plausible options, for example:

The IT capital of India is:

- A. Bangalore
- B. Mumbai
- C. Mexico
- D. Hyderabad

A well written multiple-choice question avoids obviously wrong or silly distractors (such as Mexico in the example above), so that the question makes sense when read with each of the distractors as well as with the correct answer. It is good practice to avoid "All of the above" or "None of the above" answers. If "All of the above" is used, then technically the student is correct no matter which option they select.

A more difficult and well-written multiple choice question is as follows:

Consider the following:

- I. An eight-by-eight chessboard.
- II. An eight-by-eight chessboard with two opposite corners removed.
- III. An eight-by-eight chessboard with all four corners removed.

Which of these can be tiled by two-by-one dominoes (with no overlaps or gaps, and every domino contained within the board)?

- A. I only
- B. II only
- C. I and II only
- D. I and III only
- E. I, II, and III

Advantages

There are several advantages to multiple choice tests. If item writers are well trained and items are quality assured, it can be a very effective assessment technique. If students are instructed on the way in which the item format works and myths surrounding the tests are corrected, they will perform better on the test. On many assessments, reliability has been shown to improve with larger numbers of items on a test, and with good sampling and care over case specificity, overall test reliability can be further increased.

Multiple choice tests often require less time to administer for a given amount of material than would tests requiring written responses. This results in a more comprehensive evaluation of the candidate's extent of knowledge. Even greater efficiency can be created by the use of online examination delivery software. This increase in efficiency can offset the advantages offered by free-response items. That is, if free-response items provide twice as much information but take four times as long to complete, multiple-choice items present a better measurement tool.

Multiple choice questions lend themselves to the development of objective assessment items, but without author training, questions can be subjective in nature. Because this style of test does not require a teacher to interpret answers, test-takers are graded purely on their selections, creating a lower likelihood of teacher bias in the results. Factors irrelevant to the assessed material (such as handwriting and clarity of presentation) do not come into play in a multiple-choice assessment, and so the candidate is graded purely on their knowledge of the topic. Finally, if test-takers are aware of how to use answer sheets or online examination tick boxes, their responses can be relied upon with clarity. Overall, multiple choice tests are the strongest predictors of overall student performance compared with other forms of evaluations, such as in-class participation, case exams, written assignments, and simulation games.

Disadvantages

The most serious disadvantage is the limited types of knowledge that can be assessed by multiple choice tests. Multiple choice tests are best adapted for

testing well-defined or lower-order skills. Problem-solving and higher-order reasoning skills are better assessed through short-answer and essay tests. However, multiple choice tests are often chosen, not because of the type of knowledge being assessed, but because they are more affordable for testing a large number of students. This is especially true in the United States where multiple choice tests are the preferred form of high-stakes testing.

Another disadvantage of multiple choice tests is possible ambiguity in the examinee's interpretation of the item. Failing to interpret information as the test maker intended can result in an "incorrect" response, even if the taker's response is potentially valid. The term "multiple guess" has been used to describe this scenario because test-takers may attempt to guess rather than determine the correct answer. A free response test allows the test taker to make an argument for their viewpoint and potentially receive credit.

In addition, even if students have some knowledge of a question, they receive no credit for knowing that information if they select the wrong answer and the item is scored dichotomously. However, free response questions may allow an examinee to demonstrate partial understanding of the subject and receive partial credit. Additionally if more questions on a particular subject area or topic are asked to create a larger sample then statistically their level of knowledge for that topic will be reflected more accurately in the number of correct answers and final results.

Another disadvantage of multiple choice examinations is that a student who is incapable of answering a particular question can simply select a random answer and still have a chance of receiving a mark for it. It is common practice for students with no time left to give all remaining questions random answers in the hope that they will get at least some of them right. Many exams, such as the Australian Mathematics Competition and the SAT, have systems in place to negate this, in this case by making it more beneficial to not give an answer than to give a wrong one. Another system of this is formula scoring, in which a score is proportionally reduced based on the number of incorrect responses and the number of possible choices.

In this method, the score is reduced by the number of wrong answers divided by the average number of possible answers for all questions in the test, $W/(c-1)$ where w =number of wrong responses on the test and c =the average number of possible choices for all questions on the test. All exams scored with the three-parameter model of item response theory also account for guessing. This is usually not a great issue, moreover, since the odds of a student receiving significant marks by guessing are very low when four or more selections are available.

Additionally, it is important to note that questions phrased ambiguously may cause test-taker confusion. It is generally accepted that multiple choice questions allow for only one answer, where the one answer may encapsulate a collection of previous options. However, some test creators are unaware of this and might expect the student to select multiple answers without being given explicit

permission, or providing the trailing encapsulation options. Of course, untrained test developers are a threat to validity regardless of the item format. Critics like philosopher and education proponent Jacques Derrida, said that while the demand for dispensing and checking basic knowledge is valid, there are other means to respond to this need than resorting to crib sheets. Despite being sometimes contested, the format remains popular due to its utility, reliability, and cost effectiveness.

Changing Answers

The theory that a student should trust their first instinct and stay with their initial answer on a multiple choice test is a myth. Researchers have found that although people often believe that changing answers is bad, it generally results in a higher test score. The data across twenty separate studies indicate that the percentage of “right to wrong” changes is 20.2 per cent, whereas the percentage of “wrong to right” changes is 57.8 per cent, nearly triple. Changing from “right to wrong” may be more painful and memorable (Von Restorff effect), but it is probably a good idea to change an answer after additional reflection indicates that a better choice could be made.

Alternative Response

True/False questions present candidates with a binary choice - a statement is either true or false. This method presents problems, as depending on the number of questions, a significant number of candidates could get 100 per cent just by guesswork, and should on average get 50 per cent.

Matching Type

A matching item is an item that provides a defined term and requires a test taker to match identifying characteristics to the correct term.

Completion Type

A fill-in-the-blank item provides a test taker with identifying characteristics and requires the test taker to recall the correct term. There are two types of fill-in-the-blank tests. The easier version provides a word bank of possible words that will fill in the blanks. For some exams all words in the word bank are exactly once. If a teacher wanted to create a test of medium difficulty, they would provide a test with a word bank, but some words may be used more than once and others not at all. The hardest variety of such a test is a fill-in-the-blank test in which no word bank is provided at all. This generally requires a higher level of understanding and memory than a multiple choice test. Because of this, fill-in-the-blank tests [with no word bank] are often feared by students.

Essay

Items such as short answer or essay typically require a test taker to write a response to fulfil the requirements of the item. In administrative terms, essay

items take less time to construct. As an assessment tool, essay items can test complex learning objectives as well as processes used to answer the question. The items can also provide a more realistic and generalisable task for test. Finally, these items make it difficult for test takers to guess the correct answers and require test takers to demonstrate their writing skills as well as correct spelling and grammar.

The difficulties with essay items is primarily administrative. For one, these items take more time for test takers to answer. When these questions are answered, the answers themselves are usually poorly written because test takers may not have time to organise and proofread their answers. In turn, it takes more time to score or grade these items. When these items are being scored or graded, the grading process itself becomes subjective as non-test related information may influence the process. Thus, considerable effort is required to minimize the subjectivity of the grading process. Finally, as an assessment tool, essay questions may potentially be unreliable in assessing the entire content of a subject matter.

Essay Testing

In recent times, essays have become a major part of a formal education. Secondary students are taught structured essay formats to improve their writing skills, and essays are often used by universities in selecting applicants. In both secondary and tertiary education, essays are used as testing methods to judge the mastery and comprehension of material. Students are asked to explain, comment on, or assess a topic of study in the form of an essay. Academic essays are usually more formal than literary ones. They may still allow the presentation of the writer's own views, but this is done in a logical and factual manner, with the use of the first person often discouraged.

The Five-paragraph Essay: Some students' first exposure to the genre is the five paragraph essay, a highly structured form requiring an introduction presenting the thesis statement; three body paragraphs, each of which presents an idea to support the thesis together with supporting evidence and quotations; and a conclusion, which restates the thesis and summarises the supporting points. The use of this format is controversial. Proponents argue that it teaches students how to organise their thoughts clearly in writing; opponents characterise its structure as rigid and repetitive.

A five paragraph essay usually consists of:

- The first paragraph contains the summary of topic, three supporting ideas, and the thesis.
- The second paragraph contains the first supporting idea with evidence. The last sentence of it leads into the next idea.
- The third paragraph contains the second supporting idea with the same structure as the second.
- The fourth paragraph contains the third supporting idea and the same structure as the second and third with the last sentence leading to the conclusion.

- The last paragraph restates the thesis, three supporting ideas, and gives the reader something to think about.

Academic Essays: Longer academic essays (often with a word limit of between 2,000 and 5,000 words) are often more discursive. They sometimes begin with a short summary analysis of what has previously been written on a topic, which is often called a literature review. Longer essays may also contain an introductory page in which words and phrases from the title are tightly defined. Most academic institutions will require that all substantial facts, quotations, and other supporting material used in an essay be referenced in a bibliography or works cited page at the end of the text. This scholarly convention allows others (whether teachers or fellow scholars) to understand the basis of the facts and quotations used to support the essay's argument, and thereby help to evaluate to what extent the argument is supported by evidence, and to evaluate the quality of that evidence. The academic essay tests the student's ability to present their thoughts in an organised way and tests their intellectual capabilities. Some forms of essays are:

Descriptive: Descriptive writing is characterised by sensory details, which appeal to the physical senses, and details that appeal to a reader's emotional, physical, or intellectual sensibilities. Determining the purpose, considering the audience, creating a dominant impression, using descriptive language, and organising the description are the rhetorical choices to be considered when using a description. A description is usually arranged spatially but can also be chronological or emphatic. The focus of a description is the scene. Description uses tools such as denotative language, connotative language, figurative language, metaphor, and simile to arrive at a dominant impression.

Narrative: A narrative uses tools such as flashbacks, flash-forwards, and transitions that often build to a climax. The focus of a narrative is the plot. When creating a narrative, authors must determine their purpose, consider their audience, establish their point of view, use dialogue, and organise the narrative. A narrative is usually arranged chronologically.

Exemplification: An exemplification essay is characterised by a generalisation and relevant, representative, and believable examples including anecdotes. Writers need to consider their subject, determine their purpose, consider their audience, decide on specific examples, and arrange all the parts together when writing an exemplification essay.

Comparison and Contrast: Compare and contrast is characterised by a basis for comparison, points of comparison, analogies, and either comparison by object (chunking) or by point (sequential). Comparison highlights the differences between two or more similar objects while contrasting highlights the differences between two or more objects. When writing a compare\contrast essay, writers need to determine their purpose, consider their audience, consider the basis and points of comparison, consider their thesis statement, arrange and develop the comparison, and reach a conclusion. Compare and contrast is arranged emphatically.

Cause and Effect: The defining features of a cause and effect essay are causal chains, careful language, and chronological or emphatic order. A writer using this rhetorical method must consider the subject, determine the purpose, consider the audience, think critically about different causes or consequences, consider a thesis statement, arrange the parts, consider the language, and decide on a conclusion.

Classification and Division: Classification is the categorisation of objects into a larger whole while division is the breaking of a larger whole into smaller parts.

Definition: Definition essays explain a term's meaning. Some are written about concrete terms, such as trees, oceans, and dogs, while others talk about more abstract terms, such as liberty, happiness, and virtue.

Dialectic: In this form of essay used commonly in Philosophy, one makes a thesis and argument, then objects to their own argument (with a counterargument), but then counters the counterargument with a final and novel argument. This form benefits from being more open-minded while countering a possible flaw that some may present.

Other Logical Structures: The logical progression and organisational structure of an essay can take many forms. Understanding how the movement of thought is managed through an essay has a profound impact on its overall cogency and ability to impress. A number of alternative logical structures for essays have been visualised as diagrams, making them easy to implement or adapt in the construction of an argument.

Application Essay: An admissions or application essay, sometimes also called a personal statement, is an essay or other written statement written by an applicant, often a prospective student applying to some college, university, or graduate school. The application essay is a common part of the university and college admissions process. Some applications may require one or more essays to be completed, while others make essays optional or supplementary. Essay topics range from very specific to open-ended. Common topics include career aspirations, academic strengths and weaknesses, past experiences, and reasons for applying to a particular school. The University of Chicago is known for its unusual essay prompts in its undergraduate admissions application, including “What would you do with a foot-and-a-half-tall jar of mustard”?

The Common Application, used for undergraduate admissions by many American colleges and universities, requires a general admissions essay, in addition to any supplemental admissions essays required by member institutions. The Common Application offers students six admissions essay prompts from which to choose. According to *Uni in the USA*, the Common Application essay is intended as a chance to describe “things that are unique, interesting and informative about yourself”.

The application process for All Souls College, Oxford, has the reputation of being the hardest examination in the world. It consists of several specialist papers and, until 2010, also required candidates to write an essay upon a topic

suggested by a single word such as *Possessions*, which was the topic of successful Fellow, A. L. Rowse. This was especially challenging because it provided great scope to demonstrate imagination and intelligence.

Mathematical Questions

Most mathematics questions, or calculation questions from subjects such as chemistry, physics or economics employ a style which does not fall in to any of the above categories, although some papers, notably the Maths Challenge papers in the United Kingdom employ multiple choice. Instead, most mathematics questions state a mathematical problem or exercise that requires a student to write a freehand response. Marks are given more for the steps taken than for the correct answer.

If the question has multiple parts, later parts may use answers from previous sections, and marks may be granted if an earlier incorrect answer was used but the correct method was followed, and an answer which is correct (given the incorrect input) is returned. Higher level mathematical papers may include variations on true/false, where the candidate is given a statement and asked to verify its validity by direct proof or stating a counterexample.

TEACHER-MADE TESTS

Even though parents and the media value published test scores, most teachers do not rely on standardised tests to tell them what their students know and don't know. Standardised tests occur so infrequently that one aggregate score is not very helpful in determining future instructional goals. Teacher-made tests, however, allow teachers to make decisions that keep instruction moving. Teachers can make changes immediately to meet the needs of their students.

The key to teacher-made tests is to make them a part of instruction—not separate from it. Tests should be instructional and ongoing. Rather than being “after-the-fact” to find out what students did not learn, they should be more “before-the-fact” to target essential standards. Teachers also need to make adjustments in their tests for the various learning styles, multiple intelligences and learning problems of the students in their classes. It would be impossible to address every student's needs on every test, but efforts should be made to construct tests that motivate students to learn, provide choices and make allowances for individual differences.

How can we Design Better Teacher-made Tests

Most teachers will not have time to rewrite all their tests to conform to the guidelines suggested below. However, it is important to make sure new tests are designed to meet student needs—and truly reflect learning.

PSYCHOLOGICAL TESTING

Psychological testing, also called psychometrics, the systematic use of tests to quantify psychophysical behaviour, abilities, and problems and to make

predictions about psychological performance. The word “test” refers to any means (often formally contrived) used to elicit responses to which human behaviour in other contexts can be related. When intended to predict relatively distant future behaviour (*e.g.*, success in school), such a device is called an aptitude test. When used to evaluate the individual’s present academic or vocational skill, it may be called an achievement test.

In such settings as guidance offices, mental-health clinics, and psychiatric hospitals, tests of ability and personality may be helpful in the diagnosis and detection of troublesome behaviour. Industry and government alike have been prodigious users of tests for selecting workers. Research workers often rely on tests to translate theoretical concepts (*e.g.*, intelligence) into experimentally useful measures.

General Problems of Measurement in Psychology

Physical things are perceived through their properties or attributes. A mother may directly sense the property called temperature by feeling her infant’s forehead. Yet she cannot directly observe colicky feelings or share the infant’s personal experience of hunger. She must infer such unobservable private sensations from hearing her baby cry or gurgle; from seeing him flail his arms, or frown, or smile. In the same way, much of what is called measurement must be made by inference. Thus, a mother suspecting her child is feverish may use a thermometer, in which case she ascertains his temperature by looking at the thermometer, rather than by directly touching his head.

Indeed, measurement by inference is particularly characteristic of psychology. Such abstract properties or attributes as intelligence or introversion never are directly measured but must be inferred from observable behaviour. The inference may be fairly direct or quite indirect. If persons respond intelligently (*e.g.*, by reasoning correctly) on an ability test, it can be safely inferred that they possess intelligence to some degree. In contrast, people’s capacity to make associations or connections, especially unusual ones, between things or ideas presented in a test can be used as the basis for inferring creativity, although producing a creative product requires other attributes, including motivation, opportunity, and technical skill.

Types of Measurement Scales

To measure any property or activity is to assign it a unique position along a numerical scale. When numbers are used merely to identify individuals or classes (as on the backs of athletes on a football team), they constitute a nominal scale. When a set of numbers reflects only the relative order of things (*e.g.*, pleasantness-unpleasantness of odours), it constitutes an ordinal scale. An interval scale has equal units and an arbitrarily assigned zero point; one such scale, for example, is the Fahrenheit temperature scale. Ratio scales not only provide equal units but also have absolute zero points; examples include measures of weight and distance.

Although there have been ingenious attempts to establish psychological scales with absolute zero points, psychologists usually are content with approximations to interval scales; ordinal scales often are used as well.

Primary Characteristics of Methods or Instruments

The primary requirement of a test is validity—traditionally defined as the degree to which a test actually measures whatever it purports to measure. A test is reliable to the extent that it measures consistently, but reliability is of no consequence if a test lacks validity. Since the person who draws inferences from a test must determine how well it serves his purposes, the estimation of validity inescapably requires judgement. Depending on the criteria of judgement employed, tests exhibit a number of different kinds of validity.

Empirical validity (also called statistical or predictive validity) describes how closely scores on a test correspond (correlate) with behaviour as measured in other contexts. Students' scores on a test of academic aptitude, for example, may be compared with their school grades (a commonly used criterion). To the degree that the two measures statistically correspond, the test empirically predicts the criterion of performance in school. Predictive validity has its most important application in aptitude testing (*e.g.*, in screening applicants for work, in academic placement, in assigning military personnel to different duties).

Alternatively, a test may be inspected simply to see if its content seems appropriate to its intended purpose. Such content validation is widely employed in measuring academic achievement but with recognition of the inevitable role of judgement. Thus, a geometry test exhibits content (or curricular) validity when experts (*e.g.*, teachers) believe that it adequately samples the school curriculum for that topic. Interpreted broadly, content covers desired skills (such as computational ability) as well as points of information in the case of achievement tests. Face validity (a crude kind of content validity) reflects the acceptability of a test to such people as students, parents, employers, and government officials. A test that looks valid is desirable, but face validity without some more basic validity is nothing more than window dressing.

In personality testing, judgements of test content tend to be especially untrustworthy, and dependable external criteria are rare. One may, for example, assume that a man who perspires excessively feels anxious. Yet his feelings of anxiety, if any, are not directly observable. Any assumed trait (anxiety, for example) that is held to underlie observable behaviour is called a construct. Since the construct itself is not directly measurable, the adequacy of any test as a measure of anxiety can be gauged only indirectly; *e.g.*, through evidence for its construct validity.

A test exhibits construct validity when low scorers and high scorers are found to respond differently to everyday experiences or to experimental procedures. A test presumed to measure anxiety, for example, would give evidence of construct validity if those with high scores ("high anxiety") can be shown to learn less efficiently than do those with lower scores. The rationale is that there

are several propositions associated with the concept of anxiety: anxious people are likely to learn less efficiently, especially if uncertain about their capacity to learn; they are likely to overlook things they should attend to in carrying out a task; they are apt to be under strain and hence feel fatigued. (But anxious people may be young or old, intelligent or unintelligent.) If people with high scores on a test of anxiety show such proposed signs of anxiety, that is, if a test of anxiety has the expected relationships with other measurements as given in these propositions, the test is viewed as having construct validity.

Test reliability is affected by scoring accuracy, adequacy of content sampling, and the stability of the trait being measured. Scorer reliability refers to the consistency with which different people who score the same test agree. For a test with a definite answer key, scorer reliability is of negligible concern. When the subject responds with his own words, handwriting, and organisation of subject matter, however, the preconceptions of different raters produce different scores for the same test from one rater to another; that is, the test shows scorer (or rater) unreliability. In the absence of an objective scoring key, a scorer's evaluation may differ from one time to another and from those of equally respected evaluators. Other things being equal, tests that permit objective scoring are preferred.

Reliability also depends on the representativeness with which tests sample the content to be tested. If scores on items of a test that sample a particular universe of content designed to be reasonably homogeneous (*e.g.*, vocabulary) correlate highly with those on another set of items selected from the same universe of content, the test has high content reliability. But if the universe of content is highly diverse in that it samples different factors (say, verbal reasoning and facility with numbers), the test may have high content reliability but low internal consistency.

For most purposes, the performance of a subject on the same test from day to day should be consistent. When such scores do tend to remain stable over time, the test exhibits temporal reliability. Fluctuations of scores may arise from instability of a trait; for example, the test taker may be happier one day than the next. Or temporal unreliability may reflect injudicious test construction.

Included among the major methods through which test reliability estimates are made is the comparable-forms technique, in which the scores of a group of people on one form of a test are compared with the scores they earn on another form. Theoretically, the comparable-forms approach may reflect scorer, content, and temporal reliability. This ideally demands that each form of the test be constructed by different but equally competent persons and that the forms be given at different times and evaluated by a second rater (unless an objective key is fixed).

In the test-retest method, scores of the same group of people from two administrations of the same test are correlated. If the time interval between administrations is too short, memory may unduly enhance the correlation. Or some people, for example, may look up words they missed on the first

administration of a vocabulary test and thus be able to raise their scores the second time around. Too long an interval can result in different effects for each person due to different rates of forgetting or learning. Except for very easy speed tests (*e.g.*, in which a person's score depends on how quickly he is able to do simple addition), this method may give misleading estimates of reliability.

Internal-consistency methods of estimating reliability require only one administration of a single form of a test. One method entails obtaining scores on separate halves of the test, usually the odd-numbered and the even-numbered items. The degree of correspondence (which is expressed numerically as a correlation coefficient) between scores on these half-tests permits estimation of the reliability of the test (at full length) by means of a statistical correction.

This is computed by the use of the Spearman-Brown prophecy formula (for estimating the increased reliability expected to result from increase in test length). More commonly used is a generalisation of this stepped-up, split-half reliability estimate, one of the Kuder-Richardson formulas. This formula provides an average of estimates that would result from all possible ways of dividing a test into halves.

Other Characteristics

A test that takes too long to administer is useless for most routine applications. What constitutes a reasonable period of testing time, however, depends in part on the decisions to be made from the test. Each test should be accompanied by a practicable and economically feasible scoring scheme, one scorable by machine or by quickly trained personnel being preferred.

A large, controversial literature has developed around response sets; *i.e.*, tendencies of subjects to respond systematically to items regardless of content. Thus, a given test taker may tend to answer questions on a personality test only in socially desirable ways or to select the first alternative of each set of multiple-choice answers or to malingering (*i.e.*, to purposely give wrong answers).

Response sets stem from the ways subjects perceive and cope with the testing situation. If they are tested unwillingly, they may respond carelessly and hastily to get through the test quickly. If they have trouble deciding how to answer an item, they may guess or, in a self-descriptive inventory, choose the "yes" alternative or the socially desirable one. They may even mentally reword the question to make it easier to answer. The quality of test scores is impaired when the purposes of the test administrator and the reactions of the subjects to being tested are not in harmony. Modern test construction seeks to reduce the undesired effects of subjects' reactions.

Types of Instruments and Methods

Psychophysical Scales and Psychometric, or Psychological, Scales

The concept of an absolute threshold (the lowest intensity at which a sensory stimulus, such as sound waves, is perceived) is traceable to the German

philosopher Johann Friedrich Herbart. The German physiologist Ernst Heinrich Weber later observed that the smallest discernible difference of intensity is proportional to the initial stimulus intensity. Weber found, for example, that, while people could just notice the difference after a slight change in the weight of a 10-gram object, they needed a larger change before they could just detect a difference from a 100-gram weight. This finding, known as Weber's law, is expressed more technically in the statement that the perceived (subjective) intensity varies mathematically as the logarithm of the physical (objective) intensity of the stimulus.

In traditional psychophysical scaling methods, a set of standard stimuli (such as weights) that can be ordered according to some physical property is related to sensory judgements made by experimental subjects. By the method of average error, for example, subjects are given a standard stimulus and then made to adjust a variable stimulus until they believe it is equal to the standard. The mean (average) of a number of judgements is obtained. This method and many variations have been used to study such experiences as visual illusions, tactual intensities, and auditory pitch. Psychological (psychometric) scaling methods are an outgrowth of the psychophysical tradition just described. Although their purpose is to locate stimuli on a linear (straight-line) scale, no quantitative physical values (*e.g.*, loudness or weight) for stimuli are involved. The linear scale may represent an individual's attitude towards a social institution, his judgement of the quality of an artistic product, the degree to which he exhibits a personality characteristic, or his preference for different foods.

Psychological scales thus are used for having a person rate his own characteristics as well as those of other individuals in terms of such attributes, for example, as leadership potential or initiative. In addition to locating individuals on a scale, psychological scaling can also be used to scale objects and various kinds of characteristics: finding where different foods fall on a group's preference scale; or determining the relative positions of various job characteristics in the view of those holding that job. Reported degrees of similarities between pairs of objects are used to identify scales or dimensions on which people perceive the objects.

The American psychologist L.L. Thurstone offered a number of theoretical-statistical contributions that are widely used as rationales for constructing psychometric scales. One scaling technique (comparative judgement) is based empirically on choices made by people between members of any series of paired stimuli. Statistical treatment to provide numerical estimates of the subjective (perceived) distances between members of every pair of stimuli yields a psychometric scale. Whether or not these computed scale values are consistent with the observed comparative judgements can be tested empirically.

Another of Thurstone's psychometric scaling techniques (equal-appearing intervals) has been widely used in attitude measurement. In this method judges sort statements reflecting such things as varying degrees of emotional intensity, for example, into what they perceive to be equally spaced categories; the average

(median) category assignments are used to define scale values numerically. Subsequent users of such a scale are scored according to the average scale values of the statements to which they subscribe. Another psychologist, Louis Guttman, developed a method that requires no prior group of judges, depends on intensive analysis of scale items, and yields comparable results.

Quite commonly used is the type of scale developed by Rensis Likert in which perhaps five choices ranging from strongly in favour to strongly opposed are provided for each statement, the alternatives being scored from one to five. A more general technique (successive intervals) does not depend on the assumption that judges perceive interval size accurately. The widely used graphic rating scale presents an arbitrary continuum with preassigned guides for the rater (*e.g.*, adjectives such as superior, average, and inferior).

Tests Versus Inventories

The term “test” most frequently refers to devices for measuring abilities or qualities for which there are authoritative right and wrong answers. Such a test may be contrasted with a personality inventory, for which it is often claimed that there are no right or wrong answers. At any rate, in taking what often is called a test, the subjects are instructed to do their best; in completing an inventory, they are instructed to represent their typical reactions. A distinction also has been made that in responding to an inventory the subjects control the appraisal, whereas in a test they do not. If a test is more broadly regarded as a set of stimulus situations that elicit responses from which inferences can be drawn, however, then an inventory is, according to this definition, a variety of test.

Free-response Versus Limited-response Tests

Free-response tests entail few restraints on the form or content of response, whereas limited-response tests restrict responses to one of a smaller number presented (*e.g.*, true-false). An essay test tends towards one extreme (free response), while a so-called fully objective test is at the other extreme (limited response).

Response to an essay question is not completely unlimited, however, since the answer should bear on the question. The free-response test does give practice in writing, and, when an evaluator is proficient in judging written expression, his comments on the test may aid the individual to improve his writing style. All too often, however, writing ability unfortunately affects the evaluator’s judgement of how well the test taker understands content, and this tends to reduce test reliability.

Another source of unreliability for essay tests is found in their limited sampling of content, as contrasted with the broader coverage that is possible with objective tests. Often both the scorer and the content reliability of essay tests can be improved, but such attempts are costly.

The objective test, which minimizes scorer unreliability, is best typified by the multiple-choice form, in which the subject is required to select one from

two or (preferably) more responses to a test item. Matching items that have a common set of alternatives are of this form. The true-false test question is a special multiple-choice form that may tend to arouse antagonism because of variable standards of truth or falsity.

The more general multiple-choice item is more acceptable when it is specified only that the best answer be selected; it is flexible, has high scorer reliability, and is not limited to simple factual knowledge. The ingenious test constructor can use multiple-choice items to test such functions as generalisation, application of principles, and the ability to reduce unfamiliar relationships.

Some personality tests are presented in a forced-choice format. They may, for example, force the person to choose one of two favourable words or phrases (*e.g.*, intelligent-handsome) as more descriptive of himself or one of two unfavourable terms as less descriptive (*e.g.*, stupid-ugly). Marking one choice yields a gain in score on some trait but may also preclude credit on another trait. This technique is intended to eliminate any effects from subjects' attempts to present themselves in a socially desirable light; it is not fully successful, however, because what is highly desirable for one person may be less desirable for another.

The forced-choice technique for self-appraisals is exemplified in a widely used interest inventory. Forced-choice ratings were introduced for evaluation of one military officer by another during World War II. They were an effort to avoid the preponderance of high ratings typically obtained with ordinary rating scales. Raters tend to give those being rated the benefit of any doubt, especially when they are fellow workers. Also, supervisors or teachers may give unduly favourable ratings because they believe good performance of subordinates or students reflects well on themselves.

Falling between free- and limited-response tests is a type that requires a short answer, perhaps a single word or a number, for each item. When the required response is to fit into a blank in a sentence, the test is called a completion test. This type of test is susceptible to scorer unreliability.

A personality test to which a subject responds by interpreting a picture or by telling a story it suggests resembles an essay test except that responses ordinarily are oral. A personality inventory that requires the subject to indicate whether or not a descriptive phrase applies to him is of the limited-response type. A sentence-completion personality test that asks the subject to complete statements such as "I worry because..." is akin to the short-answer and completion types.

Verbal Versus Performance Tests

A verbal (or symbol) test poses questions to which the subject supplies symbolic answers (in words or in other symbols, such as numbers). In performance tests, the subject actually executes some motor activity; for example, he assembles mechanical objects. Either the quality of performance as it takes place or its results may be rated.

The verbal test, permitting group administration, requiring no special equipment, and often being scorable by relatively unskilled evaluators, tends to

be more practical than the performance test. Both types of devices also have counterparts in personality measurement, in which verbal tests as well as behaviour ratings are used.

Written (Group) Versus Oral (Individual) Tests

The oral test is administered to one person at a time, but written tests can be given simultaneously to a number of subjects. Oral tests of achievement, being uneconomical and prone to content and scorer unreliability, have been supplanted by written tests; notable exceptions include the testing of illiterates and the anachronistic oral examinations to which candidates for graduate degrees are liable.

Proponents of individually administered intelligence tests (*e.g.*, the Stanford-Binet) state that such face-to-face testing optimises rapport and motivation, even among literate adult subjects. Oral tests of general aptitude remain popular, though numerous written group tests have been designed for the same purpose.

The interview may provide a personality measurement and, especially when it is standardised as to wording and order of questions and with a key for coding answers, may amount to an individual oral test. Used in public opinion surveys, such interviews are carefully designed to avoid the effects of interviewer bias and to be comprehensible to a highly heterogeneous sample of respondents.

Appraisal by others Versus Self-appraisal

In responding to personality inventories and rating scales, a person presumably reveals what he thinks he is like; that is, he appraises himself. Other instruments may reflect what one person thinks of another. Because self-appraisal often lacks objectivity, appraisal by another individual is common in such things as ratings for promotions. Ordinary tests of ability clearly involve evaluation of one person by another, although the subject's self-evaluation may intrude; for example, he may lack confidence to the point where he does not try to do his best.

Projective Tests

The stimuli (*e.g.*, inkblots) in a projective test are intentionally made ambiguous and open to different interpretations in the expectation that each subject will project his own unique (idiosyncratic) reactions in his answers. Techniques for evaluating such responses range from the intuitive impressions of the rater to complex, coded schemes for scoring and interpretation that require extensive manuals; some projective tests are objectively scorable.

Speed Tests Versus Power Tests

A pure speed test is homogeneous in content (*e.g.*, a simple clerical checking test), the tasks being so easy that with unlimited time all but the most incompetent of subjects could deal with them successfully. The time allowed for testing is so short, however, that even the ablest subject is not expected to finish. A useful

score is the number of correct answers made in a fixed time. In contrast, a power test (*e.g.*, a general vocabulary test) contains items that vary in difficulty to the point that no subject is expected to get all items right even with unlimited time. In practice, a definite but ample time is set for power tests.

Speed tests are suitable for testing visual perception, numerical facility, and other abilities related to vocational success. Tests of psychomotor abilities (*e.g.*, eye–hand coordination) often involve speed. Power tests tend to be more relevant to such purposes as the evaluation of academic achievement, for which the highest level of difficulty at which a person can succeed is of greater interest than his speed on easy tasks. In general, tests reflect unknown combinations of the effects of speed and power; many consist of items that vary considerably in difficulty, and the time allowed is too limited to allow a large proportion of subjects to attempt all items.

Teacher-made Versus Standardised Tests

A distinction between teacher-made tests and standardised tests is often made in relation to tests used to assess academic achievement. Ordinarily, teachers do not attempt to construct tests of general or special aptitude or of personality traits. Teacher-made tests tend instead to be geared to narrow segments of curricular content (*e.g.*, a sixth-grade geography test). Standardised tests with carefully defined procedures for administration and scoring to ensure uniformity can achieve broader goals. General principles of test construction and such considerations as reliability and validity apply to both types of test.

5

Special Measurement Techniques

Sociodrama and psychodrama were originally developed as psychotherapeutic techniques. In sociodrama, group members participate in unrehearsed drama to illuminate a general problem. Psychodrama centres on one individual in the group whose unique personal problem provides the theme. Related research techniques (*e.g.*, the sociometric test) can offer insight into interpersonal relationships. Individuals may be asked to specify members of a group whom they prefer as leader, playmate, or coworker. The choices made can then be charted in a sociogram, from which cliques or socially isolated individuals may be identified at a glance.

Research psychologists have grasped the sociometric approach as a means of measuring group cohesiveness and studying individual reactions to groups. The degree to which any group member chooses or is chosen beyond chance expectation may be calculated, and mathematical techniques may be used to determine the complex links among group members. Sociogram-choice scores have been useful in predicting such criteria as individual productivity in factory work and combat effectiveness.

Development of Standardised Tests

Test Content: Item Development Once the need for a test has been established, a plan to define its content may be prepared. For achievement tests, the test plan may also indicate thinking skills to be evaluated. Detailed content headings can be immediately suggestive of test items. It is helpful if the plan specifies weights to be allotted to different topics, as well as the desired average score and the spread of item difficulties. Whether or not such an outline is made, the

test constructor clearly must understand the purpose of the test, the universe of content to be sampled, and the forms of the items to be used.

Tryouts and Item Analysis: A set of test questions is first administered to a small group of people deemed to be representative of the population for which the final test is intended. The trial run is planned to provide a check on instructions for administering and taking the test and for intended time allowances, and it can also reveal ambiguities in the test content. After adjustments, surviving items are administered to a larger, ostensibly representative group. The resulting data permit computation of a difficulty index for each item (often taken as the percentage of the subjects who respond correctly) and of an item-test or item-subtest discrimination index (*e.g.*, a coefficient of correlation specifying the relationship of each item with total test score or subtest score).

If it is feasible to do so, measures of the relation of each item to independent criteria (*e.g.*, grades earned in school) are obtained to provide item validation. Items that are too easy or too difficult are discarded; those within a desired range of difficulty are identified. If internal consistency is sought, items that are found to be unrelated to either a total score or an appropriate subtest score are ruled out, and items that are related to available external criterion measures are identified. Those items that show the most efficiency in predicting an external criterion (highest validity) usually are preferred over those that contribute only to internal consistency (reliability).

Estimates of reliability for the entire set of items, as well as for those to be retained, commonly are calculated. If the reliability estimate is deemed to be too low, items may be added. Each alternative in multiple-choice items also may be examined statistically. Weak incorrect alternatives can be replaced, and those that are unduly attractive to higher scoring subjects may be modified.

Cross Validation: Item-selection procedures are subject to chance errors in sampling test subjects, and statistical values obtained in pretesting are usually checked (cross validated) with one or more additional samples of subjects. Typically, it is found that cross-validation values tend to shrink for many of the items that emerged as best in the original data, and further items may be found to warrant discard. Measures of correlation between total test score and scores from other, better known tests are often sought by test users.

Differential weighting: Some test items may appear to deserve extra, positive weight; some answers in multiple-choice items, though keyed as wrong, seem better than others in that they attract people who earn high scores generally. The bulk of theoretical logic and empirical evidence, nonetheless, suggests that unit weights for selected items and zero weights for discarded items and dichotomous (right versus wrong) scoring for multiple-choice items serve almost as effectively as more complicated scoring. Painstaking efforts to weight items generally are not worth the trouble.

Negative weight for wrong answers is usually avoided as presenting undue complication. In multiple-choice items, the number of answers a subject knows,

in contrast to the number he gets right (which will include some lucky guesses), can be estimated by formula. But such an average correction overanalysis the unlucky and under penalises the lucky. If the instruction is not to guess, it is variously interpreted by persons of different temperament; those who decide to guess despite the ban are often helped by partial knowledge and tend to do better.

A responsible tactic is to try to reduce these differences by directing subjects to respond to every question, even if they must guess. Such instructions, however, are inappropriate for some competitive speed tests, since candidates who mark items very rapidly and with no attention to accuracy excel if speed is the only basis for scoring; that is, if wrong answers are not penalised.

Test Norms: Test norms consist of data that make it possible to determine the relative standing of an individual who has taken a test. By itself, a subject's raw score (*e.g.*, the number of answers that agree with the scoring key) has little meaning. Almost always, a test score must be interpreted as indicating the subject's position relative to others in some group. Norms provide a basis for comparing the individual with a group.

Numerical values called centiles (or percentiles) serve as the basis for one widely applicable system of norms. From a distribution of a group's raw scores the percentage of subjects falling below any given raw score can be found. Any raw score can then be interpreted relative to the performance of the reference (or normative) group—eighth-graders, five-year-olds, institutional inmates, job applicants. The centile rank corresponding to each raw score, therefore, shows the percentage of subjects who scored below that point. Thus, 25 percent of the normative group earn scores lower than the 25th centile; and an average called the median corresponds to the 50th centile.

Another class of norm system (standard scores) is based on how far each raw score falls above or below an average score, the arithmetic mean. One resulting type of standard score, symbolised as z , is positive (*e.g.*, +1.69 or +2.43) for a raw score above the mean and negative for a raw score below the mean. Negative and fractional values can, however, be avoided in practice by using other types of standard scores obtained by multiplying z scores by an arbitrarily selected constant (say, 10) and by adding another constant (say, 50, which changes the z score mean of zero to a new mean of 50). Such changes of constants do not alter the essential characteristics of the underlying set of z scores.

The French psychologist Alfred Binet, in pioneering the development of tests of intelligence, listed test items along a normative scale on the basis of the chronological age (actual age in years and months) of groups of children that passed them. A mental-age score (*e.g.*, seven) was assigned to each subject, indicating the chronological age (*e.g.*, seven years old) in the reference sample for which his raw score was the mean. But mental age is not a direct index of brightness; a mental age of seven in a 10-year-old is different from the same mental age in a four-year-old.

To correct for this, a later development was a form of IQ (intelligence quotient), computed as the ratio of the subject's mental age to his chronological age, multiplied by 100. (Thus, the IQ made it easy to tell if a child was bright or dull for his age.)

Ratio IQs for younger age groups exhibit means close to 100 and spreads of roughly 45 points above and below 100. The classical ratio IQ has been largely supplanted by the deviation IQ, mainly because the spread around the average has not been uniform due to different ranges of item difficulty at different age levels. The deviation IQ, a type of standard score, has a mean of 100 and a standard deviation of 16 for each age level. Practice with the Stanford-Binet test reflects the finding that average performance on the test does not increase beyond age 18. Therefore, the chronological age of any individual older than 18 is taken as 18 for the purpose of determining IQ.

The Stanford-Binet has been largely supplanted by several tests developed by the American psychologist David Wechsler between the late 1930s and the early 1960s. These tests have subtests for several capacities, some verbal and some operational, each subtest having its own norms. After constructing tests for adults, Wechsler developed tests for older and for younger children.

Assessing test structure: Factor analysis Factor analysis is a method of assessment frequently used for the systematic analysis of intellectual ability and other test domains, such as personality measures. Just after the turn of the 20th century the British psychologist Charles E. Spearman systematically explored positive intercorrelations between measures of apparently different abilities to provide evidence that much of the variability in scores that children earn on tests of intelligence depends on one general underlying factor, which he called *g*. In addition he believed that each test contained an *s* factor specific to it alone.

In the United States, Thurstone developed a statistical technique called multiple-factor analysis, with which he was able to demonstrate, in a set of tests of intelligence, that there were primary mental abilities, such as verbal comprehension, numerical computation, spatial orientation, and general reasoning. Although later work has supported the differentiation between these abilities, no definitive taxonomy of abilities has become established. One element in the problem is the finding that each such ability can be shown to be composed of narrower factors.

The first computational methods in factor analysis have been supplanted by mathematically more elegant, computer-generated solutions. While earlier techniques were primarily exploratory, the Swedish statistician Karl Gustav Jöreskog and others have developed procedures that permit the researcher to test hypotheses about the structure in a set of data.

Rooted in extensive applications of factor analysis, a structure-of-intellect model developed by the American psychologist Joy Paul Guilford posited a very large number of factors of intelligence. Guilford envisaged three intersecting dimensions corresponding respectively to four kinds of test content, five kinds

of intellectual operation, and six kinds of product. Each of the 120 cells in the cube thus generated was hypothesised to represent a separate ability, each constituting a distinct factor of intellect. Educational and vocational counsellors usually prefer a substantially smaller number of scores than the 120 implied by this model. Factor analysis has also been widely used outside the realm of intelligence, especially to seek the structure of personality as reflected in ratings by oneself and by others. Although there is even less consensus here than for intelligence, a number of studies suggest that four prevalent factors can be approximately labelled, namely, conformity, extroversion, anxiety, and dependability.

Profile analysis: With the fractionation of tests (*e.g.*, to yield scores measuring separate factors or clusters), new concern has arisen for interpreting differences among scores measuring the underlying variables, however conceived. Scores of an individual on several such measures can be plotted graphically as a profile; for direct comparability, all raw scores may be expressed in terms of standard scores that have equal means and variabilities. The difference between any pair of scores that have less than perfect reliability tends to be less reliable than either, and fluctuations in the graph should be interpreted cautiously. Nevertheless, various features of an individual's profile may be examined, such as scatter (fluctuation from one measure to another) and relative level of performance on different measures.

(The particular shape of the graph, it should be noted, partly depends upon the arbitrary order in which measures are listed.) One may also statistically express the degree of similarity between any two profiles. Such statistical measures of pattern similarity permit quantitative comparison of profiles for different persons, of profiles of the same individual's performance at different times, of individual with group profiles, or of one group profile with another. Comparison of an individual's profile with similar graphs representing the means for various occupational groups, for example, is useful for vocational guidance or personnel selection.

Teacher-Made Assessment

For an assessment to be high quality it needs to have good validity and reliability as well as absence from bias.

Validity: is the evaluation of the "adequacy and appropriateness of the interpretations and uses of assessment results" for a given group of individuals (Linn & Miller, 2005, p. 68). For example, is it appropriate to conclude that the results of a mathematics test on fractions given to recent immigrants accurately represents their understanding of fractions? Is it appropriate for the teacher to conclude, based on her observations, that a kindergarten student, Jasmine, has Attention Deficit Disorder because she does not follow the teachers oral instructions? Obviously in each situation other interpretations are possible – that the immigrant students have poor English skills rather than mathematics skills, or that Jasmine may be hearing impaired.

It is important to understand that validity refers to the *interpretation* and *uses* made of the results of an assessment procedure, not of the assessment procedure itself. For example, making judgements about the results of the same test on fractions may be valid if the students all understand English well. A teacher concluding from her observations that the kindergarten student has Attention Deficit Disorder may be appropriate if the student has been screened for hearing and other disorders (although the classification of a disorder like ADD cannot be made by one teacher). Validity involves making an overall judgement of the degree to which the interpretations and uses of the assessment results are justified. Validity is a matter of degree (*e.g.*, high, moderate, or low validity) rather than all-or none (*e.g.*, totally valid vs. invalid).

Three sources of evidence are considered when assessing validity - content, construct and predictive. Content validity is associated with the question: How well does the assessment include the content or tasks it is supposed to? For example, suppose your educational psychology instructor devises a mid-term test and tells you this includes chapters one to seven in the text book. Obviously, all the items in test should be based on the content from educational psychology, not your methods or cultural foundations classes. Also, the items in the test should cover content from all seven chapters and not just chapters three to seven - unless the instructor tells you that these chapters have priority.

Teachers' have to be clear about their purposes and priorities for instruction before they can begin to gather evidence related content validity. Content validation determines the degree that assessment tasks are relevant and representative of the tasks judged by the teacher (or test developer) to represent their goals and objectives. It is important for teachers to think about content validation when devising assessment tasks and one way to help do this is to devise a Table of Specifications. An example, based on Pennsylvania's State standards for grade 3 geography. In the left hand column is the instructional content for a 20-item test the teacher has decided to construct with two kinds of instructional objectives: identification and uses or locates.

The second and third columns identify the number of items for each content area and each instructional objective. Notice that the teacher has decided that six items should be devoted to the sub area of geographic representations- more than any other sub area. Devising a table of specifications helps teachers determine if some content areas or concepts are over-sampled (*i.e.*, there are too many items) and some concepts are under-sampled (*i.e.*, there are too few items).

Construct validity is more complex than content validity evidence. Often we are interested in making broader judgements about student's performances than specific skills such as doing fractions. The focus may be on constructs such as mathematical reasoning or reading comprehension. A construct is a characteristic of a person we assume exists to help explain behaviour. For example, we use the concept of test anxiety to explain why some individuals when taking a test have difficulty concentrating, have physiological reactions such as sweating,

and perform poorly on tests but not in class assignments. Similarly mathematics reasoning and reading comprehension are constructs as we use them to help explain performance on an assessment.

Construct validation is the process of determining the extent to which performance on an assessment can be interpreted in terms of the intended constructs and is not influenced by factors irrelevant to the construct. For example, judgements about recent immigrants' performance on a mathematical reasoning test administered in English will have low construct validity if the results are influenced by English Language skills that are irrelevant to mathematical problem solving. Similarly, construct validity of end-of-semester examinations is likely to be poor for those students who are highly anxious when taking major tests but not during regular class periods or when doing assignments. Teachers can help increase construct validity by trying to reduce factors that influence performance but are irrelevant to the construct being assessed. These factors include anxiety, English language skills, and reading speed.

A third form of validity evidence is called criterion-related validity. Selective colleges use the ACT or SAT among other criteria to choose who will be admitted because these standardised tests help predict freshman grades, *i.e.*, have high criterion-related validity. Some K-12 schools give students math or reading tests in Fall semester in order to predict which are likely to do well on the annual State Tests administered in Spring semester and which students are unlikely to pass the tests and so will need additional assistance. If the tests administered in Fall do not predict students' performances accurately then the additional assistance may be given to the wrong students illustrating the importance of criterion-related validity.

Reliability: Refers to the *consistency* of a measurement (Linn & Miller 2005). Suppose Mr. Garcia is teaching a unit on food chemistry in his 10th grade class and gives an assessment at the end of the unit using test items from the teachers' guide. Reliability is related to questions such as: How similar would the scores of the students be if they had taken the assessment on a Friday or Monday?

Would the scores have varied if Mr. Garcia had selected different test items, or if a different teacher had graded the test? An assessment provides information about students by using a specific measure of performance at one particular time. Unless the results from the assessment are reasonably consistent over different occasions, different raters, or different tasks (in the same content domain) confidence in the results will be low and so cannot be useful in improving student learning.

Obviously we cannot expect perfect consistency. Students' memory, attention, fatigue, effort, and anxiety fluctuate and so influence performance. Even trained raters vary somewhat when grading assessment such as essays, a science project, or an oral presentation. Also, the wording and design of specific items influence students' performances.

However, some assessments are more reliable than others and there are several strategies teachers can use to increase reliability:

1. *Assessments with more tasks or items typically have higher reliability.* To understand this, consider two tests one with five items and one with 50 items. Chance factors influence the shorter test more than the longer test. If a student does not understand one of the items in the first test the total score is very highly influenced (it would be reduced by 20 per cent). In contrast, if there was one item in the test with 50 items that were confusing, the total score would be influenced much less (by only 2 per cent). Obviously this does not mean that assessments should be inordinately long, but, on average, enough tasks should be included to reduce the influence of chance variations.
2. *Clear directions and tasks help increase reliability.* If the directions or wording of specific tasks or items are unclear, then students have to guess what they mean undermining the accuracy of their results.
3. *Clear scoring criteria are crucial in ensuring high reliability.* In a later section of this chapter we describe strategies for developing scoring criteria for a variety of types of assessment.

Absence of Bias: Bias occurs in assessment when there are components in the assessment method or administration of the assessment that distort the performance of the student because of their personal characteristics such as gender, ethnicity, or social class. Two types of assessment bias are important: offensiveness and unfair penalisation. An assessment is most likely to be *offensive to a subgroup of students* when negative stereotypes are included in the test. For example, the assessment in a health class could include items in which all the doctors were men and all the nurses were women. Or a series of questions in a social studies class could portray Latinos and Asians as immigrants rather than native born Americans.

In these examples, some female, Latino or Asian students are likely to be offended by the stereotypes, and this can distract them from performing well on the assessment.

Unfair penalisation occurs when items disadvantage one group not because they may be offensive but because of differential background experiences. For example, an item for math assessment that assumes knowledge of a particular sport may disadvantage groups not as familiar with that sport (*e.g.*, American football for recent immigrants). Or an assessment on team work that asks students to model their concept of a team on a symphony orchestra is likely to be easier for those students who have attended orchestra performances - probably students from affluent families.

Unfair penalisation does not occur just because some students do poorly in class. For example, asking questions about a specific sport in a Physical Education class when information on that sport had been discussed in class is not unfair penalisation as long as the questions do not require knowledge beyond that taught in class that some groups are less likely to have.

STANDARDISED TEST

A standardised test or Examination system is a test that is administered and scored in a consistent, or “standard”, manner. Standardised tests are designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent and are administered and scored in a predetermined, standard manner. Any test in which the same test is given in the same manner to all test takers is a standardised test. Standardised tests need not be high-stakes tests, time-limited tests, or multiple-choice tests. The opposite of a standardised test is a *non-standardised test*. Non-standardised testing gives significantly different tests to different test takers, or gives the same test under significantly different conditions (*e.g.*, one group is permitted far less time to complete the test than the next group), or evaluates them differently (*e.g.*, the same answer is counted right for one student, but wrong for another student). Standardised tests are perceived as being more fair than non-standardised tests. The consistency also permits more reliable comparison of outcomes across all test takers.

History

China

The earliest evidence of standardised testing was in China, where the imperial examinations covered the Six Arts which included music, archery and horsemanship, arithmetic, writing, and knowledge of the rituals and ceremonies of both public and private parts. Later, sections on military strategies, civil law, revenue and taxation, agriculture and geography were added to the testing. In this form, the examinations were institutionalised during the 6th century CE, under the Sui Dynasty.

Britain

Standardised testing was introduced into Europe in the early 19th century, modelled on the Chinese mandarin examinations, through the advocacy of British colonial administrators, the most “persistent” of which was Britain’s consul in Guangzhou, China, Thomas Taylor Meadows. Meadows warned of the collapse of the British Empire if standardised testing was not implemented throughout the empire immediately.

Prior to their adoption, standardised testing was not traditionally a part of Western pedagogy; based on the sceptical and open-ended tradition of debate inherited from Ancient Greece, Western academia favoured non-standardised assessments using essays written by students. It is because of this that the first European implementation of standardised testing did not occur in Europe proper, but in British India. Inspired by the Chinese use of standardised testing, in the early 19th century, British “company managers hired and promoted employees based on competitive examinations in order to prevent corruption and favouritism.” This practice of standardised testing was later adopted in the late

19th century by the British mainland. The parliamentary debates that ensued made many references to the “Chinese mandarin system.” It was from Britain, that standardised testing spread, not only throughout the British Commonwealth, but to Europe and then America. Its spread was fuelled by the Industrial Revolution. Given the large number of school students during and after the Industrial Revolution, when compulsory education laws increased student populations, open-ended assessment of all students decreased. Moreover, the lack of a standardised process introduces a substantial source of measurement error, as graders might show favouritism or might disagree with each other about the relative merits of different answers.

More recently, it has been shaped in part by the ease and low cost of grading of multiple-choice tests by computer. Grading essays by computer is more difficult, but is also done. In other instances, essays and other open-ended responses are graded according to a pre-determined assessment rubric by trained graders.

United States

The use of standardised testing in the United States is a 20th-century phenomenon with its origins in World War I and the Army Alpha and Beta tests developed by Robert Yerkes and colleagues. In the United States, the need for the federal government to make meaningful comparisons across a highly decentralised (locally controlled) public education system has also contributed to the debate about standardised testing, including the Elementary and Secondary Education Act of 1965 that required standardised testing in public schools. US Public Law 107-110, known as the No Child Left Behind Act of 2002 further ties public school funding to standardised testing.

Design and Scoring

Standardised testing can be composed of multiple-choice questions, true-false questions, essay questions, authentic assessments, or nearly any other form of assessment. Multiple-choice and true-false items are often chosen because they can be given and scored inexpensively and quickly by scoring special answer sheets by computer or via computer-adaptive testing. Some standardised tests have short-answer or essay writing components that are assigned a score by independent evaluators who use rubrics (rules or guidelines) and benchmark papers (examples of papers for each possible score) to determine the grade to be given to a response. Most assessments, however, are not scored by people; people are used to score items that are not able to be scored easily by computer (*i.e.*, essays). For example, the Graduate Record Exam is a computer-adaptive assessment that requires no scoring by people (except for the writing portion).

Scoring Issues

Human scoring is often variable, which is why computer scoring is preferred when feasible. For example, some believe that poorly paid employees will score

tests badly. Agreement between scorers can vary between 60 to 85 percent, depending on the test and the scoring session. Sometimes states pay to have two or more scorers read each paper; if their scores do not agree, then the paper is passed to additional scorers.

Open-ended components of tests are often only a small proportion of the test. Most commonly, a major test includes both human-scored and computer-scored sections. These major tests do not measure the student's overall ability in learning.

Table. Sample Scoring for the History Question: What Caused World War II

Student Answers	Standardized Grading	Non-standardized Grading
	Grading rubric: Answers must be marked correct if they mention at least one of the following: Germany's the invasion of Poland, Japan's invasion of China, or economic issues.	No grading standards. Each teacher grades however he wants to, considering factors like the answer, the student's academic potential, and attitude.
<i>Student #1:</i> WWII was caused by Hitler and Germany invading Poland.	<i>Teacher #1:</i> This answer mentions one of the required items, so it is correct. <i>Teacher #2:</i> This answer is correct.	<i>Teacher #1:</i> I feel like this answer is good enough, so I'll mark it correct. <i>Teacher #2:</i> This answer is correct, but this good student should be able to do better than that, so I'll only give partial credit.
<i>Student #2:</i> WWII was caused by multiple factors, including the Great Depression and the general economic situation, the rise of nationalism, fascism, and imperialist expansionism, and unresolved resentments related to WWI. The war in Europe began with the German invasion of Poland.	<i>Teacher #1:</i> This answer mentions one of the required items, so it is correct. <i>Teacher #2:</i> This answer is correct.	<i>Teacher #1:</i> I feel like this answer is correct and complete, so I'll give full credit. <i>Teacher #2:</i> I feel like this answer is correct, so I'll give full points.
<i>Student #3:</i> WWII was caused by the assassination of Archduke Ferdinand.	<i>Teacher #1:</i> This answer does not mention any of the required items. No points. <i>Teacher #2:</i> This answer is wrong. No credit.	<i>Teacher #1:</i> This answer is wrong. No points. <i>Teacher #2:</i> This answer is wrong, but this student tried hard and the sentence is grammatically correct, so I'll give one point for effort.

There are two types of standardised test score interpretations: a norm-referenced score interpretation or a criterion-referenced score interpretation:

- Norm-referenced score interpretations compare test-takers to a sample of peers. The goal is to rank students as being better or worse than other students. Norm-referenced test score interpretations are associated with traditional education. Students who perform better than others pass the test, and students who perform worse than others fail the test.
- Criterion-referenced score interpretations compare test-takers to a criterion (a formal definition of content), regardless of the scores of other examinees. These may also be described as standards-based assessments, as they are aligned with the standards-based education reform movement. Criterion-referenced score interpretations are concerned solely with whether or not this particular student's answer is correct and complete. Under criterion-referenced systems, it is possible for all students to pass the test, or for all students to fail the test.

Either of these systems can be used in standardised testing. What is important to standardised testing is whether all students are asked equivalent questions, under equivalent circumstances, and graded equally. In a standardised test, if a given answer is correct for one student, it is correct for all students. Graders do not accept an answer as good enough for one student but reject the same answer as inadequate for another student.

Standards

The considerations of validity and reliability typically are viewed as essential elements for determining the quality of any standardised test. However, professional and practitioner associations frequently have placed these concerns within broader contexts when developing standards and making overall judgements about the quality of any standardised test as a whole within a given context.

Testing Standards

In the field of psychometrics, the *Standards for Educational and Psychological Testing* place standards about validity and reliability, along with errors of measurement and issues related to the accommodation of individuals with disabilities. The third and final major topic covers standards related to testing applications, credentialing, plus testing in programme evaluation and public policy.

Advantages

One of the main advantages of standardised testing is that the results can be empirically documented; therefore, the test scores can be shown to have a relative degree of validity and reliability, as well as results which are generalisable and replicable. This is often contrasted with grades on a school transcript, which

are assigned by individual teachers. It may be difficult to account for differences in educational culture across schools, difficulty of a given teacher's curriculum, differences in teaching style, and techniques and biases that affect grading. This makes standardised tests useful for admissions purposes in higher education, where a school is trying to compare students from across the nation or across the world.

Another advantage is aggregation. A well designed standardised test provides an assessment of an individual's mastery of a domain of knowledge or skill which at some level of aggregation will provide useful information. That is, while individual assessments may not be accurate enough for practical purposes, the mean scores of classes, schools, branches of a company, or other groups may well provide useful information because of the reduction of error accomplished by increasing the sample size. Standardised tests, which by definition give all test-takers the same test under the same (or reasonably equal) conditions, are also perceived as being more fair than assessments that use different questions or different conditions for students according to their race, socioeconomic status, or other considerations.

Disadvantages and Criticism

“Standardised tests can't measure initiative, creativity, imagination, conceptual thinking, curiosity, effort, irony, judgement, commitment, nuance, good will, ethical reflection, or a host of other valuable dispositions and attributes. What they can measure and count are isolated skills, specific facts and function, content knowledge, the least interesting and least significant aspects of learning.”

— Bill Ayers

Standardised tests are useful tools for assessing student achievement, and can be used to focus instruction on desired outcomes, such as reading and math skills. However, critics feel that overuse and misuse of these tests harms teaching and learning by narrowing the curriculum. According to the group FairTest, when standardised tests are the primary factor in accountability, schools use the tests to define curriculum and focus instruction. Critics say that “teaching to the test” disfavour higher-order learning. While it is possible to use a standardised test without letting its contents determine curriculum and instruction, frequently, what is not tested is not taught, and how the subject is tested often becomes a model for how to teach the subject.

Uncritical use of standardised test scores to evaluate teacher and school performance is inappropriate, because the students' scores are influenced by three things: what students learn in school, what students learn outside of school, and the students' innate intelligence. The school only has control over one of these three factors. Value-added modelling has been proposed to cope with this criticism by statistically controlling for innate ability and out-of-school contextual factors. In a value-added system of interpreting test scores, analysts estimate an expected score for each student, based on factors such as the student's

own previous test scores, primary language, or socioeconomic status. The difference between the student's expected score and actual score is presumed to be due primarily to the teacher's efforts.

Supporters of standardised testing respond that these are not reasons to abandon standardised testing in favour of either non-standardised testing or of no assessment at all, but rather criticisms of poorly designed testing regimes. They argue that testing does and should focus educational resources on the most important aspects of education — imparting a pre-defined set of knowledge and skills — and that other aspects are either less important, or should be added to the testing scheme.

In her book, *Now You See It*, Cathy Davidson criticizes standardised tests. She describes our youth as “assembly line kids on an assembly line model,” meaning the use of standardised test as a part of a one-size-fits-all educational model. She also criticizes the narrowness of skills being tested and labelling children without these skills as failures or as students with disabilities. Widespread and organised cheating has been a growing culture in today's reformation of schools.

Scoring Information Loss

A test question might require a student to calculate the area of a triangle. Compare the information provided in these two answers.

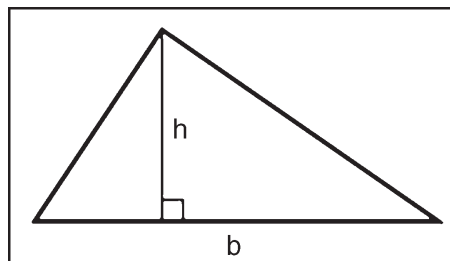
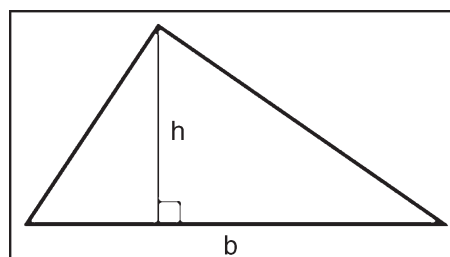


Fig. Area = 7.5 cm².



Base = 5 cm; Height = 3 cm

Area = $\frac{1}{2}(\text{Base} \times \text{Height})$

Area = $\frac{1}{2}(5 \text{ cm} \times 3 \text{ cm})$

Area = 7.5 cm²

The first shows scoring information loss. The teacher knows whether the student got the right answer, but does not know how the student arrived at the

answer. If the answer is wrong, the teacher does not know whether the student was guessing, made a simple error, or fundamentally misunderstands the subject.

When tests are scored *right-wrong*, an important assumption has been made about learning. The number of *right* answers or the sum of item scores (where partial credit is given) is assumed to be the appropriate and sufficient measure of current performance status. In addition, a secondary assumption is made that there is no meaningful information in the *wrong* answers.

In the first place, a correct answer can be achieved using *memorisation* without any profound understanding of the underlying content or conceptual structure of the problem posed. Second, when more than one step for solution is required, there are often a variety of approaches to answering that will lead to a *correct* result. The fact that the answer is correct does not indicate which of the several possible procedures were used. When the student supplies the answer (or shows the work) this information is readily available from the original documents.

Second, if the *wrong* answers were *blind* guesses, there would be no information to be found among these answers. On the other hand, if *wrong* answers reflect interpretation departures from the expected one, these answers should show an ordered relationship to whatever the overall test is measuring. This departure should be dependent upon the level of psycholinguistic maturity of the student choosing or giving the answer in the vernacular in which the test is written.

In this second case it should be possible to extract this order from the responses to the test items. Such extraction processes, the Rasch model for instance, are standard practice for item development among professionals. However, because the *wrong* answers are discarded during the scoring process, attempts to interpret these answers for the information they might contain is seldom undertaken.

Third, although topic-based subtest scores are sometimes provided, the more common practice is to report the total score or a rescaled version of it. This rescaling is intended to compare these scores to a standard of some sort. This further collapse of the test results systematically removes all the information about which particular items were missed.

Thus, scoring a test right-wrong loses:

1. How students achieved their *correct* answers,
2. What led them astray towards unacceptable answers and
3. Where within the body of the test this departure from expectation occurred.

This commentary suggests that the current scoring procedure conceals the dynamics of the test-taking process and obscures the capabilities of the students being assessed. Current scoring practice oversimplifies these data in the initial scoring step. The result of this procedural error is to obscure of the diagnostic information that could help teachers serve their students better. It further prevents those who are diligently preparing these tests from being able to observe the information that would otherwise have alerted them to the presence of this error.

A solution to this problem, known as Response Spectrum Evaluation (RSE), is currently being developed that appears to be capable of recovering all three of these forms of information loss, while still providing a numerical scale to establish current performance status and to track performance change. This RSE approach provides an interpretation of the thinking processes behind every answer (both the right and the wrong ones) that tells teachers how they were thinking for every answer they provide. Among other findings, this chapter reports that the recoverable information explains between two and three times more of the test variability than considering only the right answers.

This massive loss of information can be explained by the fact that the “wrong” answers are removed from the test information being collected during the scoring process and is no longer available to reveal the procedural error inherent in right-wrong scoring. The procedure bypasses the limitations produced by the linear dependencies inherent in test data.

Testing bias occurs when a test systematically favours one group over another, even though both groups are equal on the trait the test measures. Critics allege that test makers and facilitators tend to represent a middle class, white background. Critics claim that standardised testing match the values, habits, and language of the test makers. However, being that most tests come from a white, middle-class background, it is important to note that the highest scoring groups are not people of that background, but rather tend to come from Asian populations.

Not all tests are well-written, for example, containing multiple-choice questions with ambiguous answers, or poor coverage of the desired curriculum. Some standardised tests include essay questions, and some have criticized the effectiveness of the grading methods. Recently, partial computerised grading of essays has been introduced for some tests, which is even more controversial.

Educational Decisions

Test scores are in some cases used as a sole, mandatory, or primary criterion for admissions or certification. For example, some U.S. states require high school graduation examinations. Adequate scores on these exit exams are required for high school graduation. The General Educational Development test is often used as an alternative to a high school diploma.

Other applications include tracking (deciding whether a student should be enrolled in the “fast” or “slow” version of a course) and awarding scholarships. In the United States, many colleges and universities automatically translate scores on Advanced Placement tests into college credit, satisfaction of graduation requirements, or placement in more advanced courses.

Generalised tests such as the SAT or GRE are more often used as one measure among several, when making admissions decisions. Some public institutions have cutoff scores for the SAT, GPA, or class rank, for creating classes of applicants to automatically accept or reject. Heavy reliance on standardised tests for decision-making is often controversial, for the reasons noted above.

Critics often propose emphasizing cumulative or even non-numerical measures, such as classroom grades or brief individual assessments (written in prose) from teachers. Supporters argue that test scores provide a clear-cut, objective standard that minimizes the potential for political influence or favouritism.

The National Academy of Sciences recommends that major educational decisions not be based solely on a test score. The use of minimum cut-scores for entrance or graduation does not imply a single standard, since test scores are nearly always combined with other minimal criteria such as number of credits, prerequisite courses, attendance, *etc.* Test scores are often perceived as the “sole criteria” simply because they are the most difficult, or the fulfilment of other criteria is automatically assumed. One exception to this rule is the GED, which has allowed many people to have their skills recognised even though they did not meet traditional criteria. *The Role and Importance of Standardised Testing in the World of Teaching and Training*

Standardised testing has been called the greatest single social contribution of modern psychology, and it may be the most useful evaluation method available for human resource-intensive endeavours. For most of their history, however, standardised tests have been developed and administered on a large scale and large, typically politically-sensitive organisations have controlled their use.

In the United States, standardised tests’ political exposure has sometimes compromised their use despite the intrepid efforts of psychometricians to maintain their integrity. Some of you may recall the infamous “Lake Wobegon” scandal of the 1980s when a medical doctor, John J. Cannell, discovered that every U.S. state claimed an average student score on nationally-normed tests that was above the national average (Phelps, 2005b). Less well known, perhaps, are the persistent efforts of many powerful groups of professional educators to either eliminate the use of standardised tests or limit their use to the most unreliable types (Phelps, 2003).

With powerful forces opposed to the use (or to the proper use) of a beneficial technology that is typically provided by large, politically-sensitive organisations, perhaps it is time to consider alternative methods of providing that beneficial technology. One such alternative method is the topic of today’s session.

Why Standardised Testing?

Standardised tests are not perfect evaluation tools. Used validly and reliably, however, standardised tests provide decision-makers useful information that no other evaluation method can provide. Many research studies on educational testing dating back to the early part of the 19th century have compared different teachers’ evaluations of identical student work or compared the consistency of teachers’ marks to those of standardised test results over time. Not surprisingly, researchers found wide variance from teacher to teacher in grading identical student work or over time with the same teacher.

In the 1910s, for example, researchers Starch and Elliott (1912) made copies of two actual English examination papers and sent them to teachers to grade

and return. The marks ranged from 50 to 98 percent. One paper, graded by 142 teachers, received fourteen marks below 80 percent and fourteen above 94 percent. "That is, a paper which was considered too poor for a passing grade by some teachers was rated as excellent by others."

Starch and Elliot repeated the procedure with duplicate Geometry tests (1913). Teachers' marks on the 116 returned papers ranged from 28 to 92 percent, with twenty grades below 60 percent and nine of 85 percent and above. According to Lincoln and Workman (1936, 7); This type of experiment has been repeated many times by investigators and always with similar results. Therefore there is abundant evidence that teachers' marks are a very unreliable means of measurement.

Without standardised tests (or standardised grading protocols) in education, we would increase our reliance on individual teacher grading and testing. Are teacher evaluations free of standardised testing's alleged failings? No. Individual teachers can *narrow the curriculum* to that which they prefer. Grades are susceptible to *inflation* with ordinary teachers, as students get to know a teacher better and learn his idiosyncrasies. A teacher's (or school's) grades and test scores are far less likely to be generalisable than any standardised tests'.

According to the research on the topic, many U.S teachers consider "nearly everything" when assigning marks, including student class participation, perceived effort, progress over the period of the course, and comportment, according to one researcher. Actual achievement vis-à-vis the subject matter is just one factor. One study of teacher grading practices discovered that 66 percent of teachers felt that their perception of a student's ability should be taken into consideration in awarding the final grade.

When individual teachers, or individual employers for that matter, are given the responsibility to make judgements unanchored by common standards or rules, those judgements tend to float freely in the currents of time, fitting first one context, then another, and then another. Being idiosyncratic to each particular, temporary context, each free-floating evaluation result is not generalisable to any permanent context. It is a judgement that makes sense only to a particular teacher or employer at a particular point in time and space.

In past, standardised tests were often called "objective tests," which implied that teacher-made tests were "subjective." Standardised tests' clear separation from the influence of local decision-makers, be they classroom teachers or personnel managers responsible for hiring new employees, remains one of their most beneficial features. The adoption of standardised university admission testing in the United States in the mid-twentieth century, for example, helped to pave the way for minorities who lacked the familial connections and social pedigree of wealthy WASPs (*i.e.*, White, Anglo-Saxon Protestants). According to Professor Stephen G. Sireci (2005, 113), the bad reputation of standardised tests portrayed by some critics "is an undeserved one." He continues

People accuse standardised tests of being unfair, biased and discriminatory. Believe it or not, standardised tests are actually designed to promote test fairness.

Standardised simply means that the test content is equivalent across administrations and that the conditions under which the test is administered are the same for all test takers. ...Standardised tests are used to provide objective information. For example, employment tests are used to avoid unethical hiring practices (*e.g.*, nepotism, ethnic discrimination, *etc.*). If an assessment system uses tests that are not standardised, the system is likely to be unfair to many candidates.

There is more to subjectivity in decision-making than ethnic, racial, gender, or class bias, however. The fact is that true objectivity requires too much time to be practical in making everyday decisions. Double-blind controlled experiments or programme evaluations with random assignment require time, money, and trained professional observation to monitor their progress. In our daily lives, we make judgements and decisions continuously. We cannot set up a controlled experiment, and wait for the results, every time we must choose which laundry detergent to purchase, where to go on vacation or, for that matter, whom to hire for a job or whom to admit to the last available place at university.

The time-saving decision-making technique we typically use to get on with our lives, apparently, is Bayesian reasoning, named for the early 18th-century statistician Thomas Bayes. In Bayesian reasoning, we employ what relevant prior knowledge we have to each decision. We calculate the “subjective probabilities,” which are not, in the strictest meaning of the term really “subjective.” More accurately, they are incomplete probabilities that incorporate the information we have accumulated that is relevant to the matter at hand. That information may be reliable or not, verified or not, true or not. Nonetheless, until we discover a Fountain of Youth to provide us everlasting life, we must rely on Bayesian reasoning as a time-saving heuristic to negotiate our lives in the short time allotted to each of us (“Bayes Rules,” 2006).

Thus, a standardised test is more than an antidote to biased judgement. We need standardised tests because each of us is a prisoner of our own limited experiences and observations. Standardised tests provide an opportunity to make decisions about individuals that are free of subjectivity, be that subjectivity due to bias or Bayesian shortcuts. In developing standardised tests, trained professionals collect empirical data, apply statistical benchmarks, and make detached, objective evaluations.

Standardised Testing: The Long View

Standardised tests have provided information for making important decisions at least since the first administration of the Chinese civil service examination many centuries ago. The “scientific” standardised test (with statistically-calibrated score scales), however, is just a century old. The innovators responsible for the development of the scientific standardised test—*e.g.*, Binet, Simon, Rice, Thorndike—though, likely would be amazed by the improvements made in testing technology within the relatively brief period since—*e.g.*, computer-adaptive testing or open-source, Web-based platforms, such as the Examination Assessment

Management System (ExAMS). It would seem that testing technology has improved over time exponentially. Test developers have increased the complexity and technical sophistication of their product in response to market and regulatory demands. Today's standardised tests are better in most every way than their progenitors. They provide more information for the price, and they are more reliable, fair, and valid (when used as they are designed to be used).

But, the exponential rate of improvement carries some risk. At the same time standardised tests have improved in quality and convenience, they have become more difficult for the average person or policymaker to understand. Most standardised tests administered a century ago were simply larger-scale, standardised versions of an ordinary classroom teacher's examination. In all apparent aspects, they looked familiar to the average examinee.

Some of today's standardised tests might seem to the average citizen or policymaker as different in character from their 100-year-old ancestors as today's airplanes or automobiles do from their 100-year-old antecedents. Any of you who have tried in plain language to explain to policy makers the concepts of item response theory, differential item functioning, computer-adaptive testing, or point-biserial correlation will know what I mean.

The combination of technical complexity and the widespread use of testing for public purposes should elicit a clear, measured, and open public discussion on testing policy. And, I hope that it does where you live. In the United States, unfortunately, the public and policymakers are generally showered with obfuscation, misinformation, and disinformation.

The Testing Policy Debate in the United States: The Sound of One Hand Clapping

Standardised testing in the United States is an enigma. Arguably, the country hosts much of the world's most advanced technical research and innovation. Yet, debates on testing policy remain primitive and one-sided.

The late economist Mancur Olsen (1965, 1982) developed a theory to explain the political power of "special interests" in democratic societies. Individuals join groups that provide private benefits, such as protection against market competition, disruptive technologies, or other challenges to the familiarity and security of the *status quo* like those portended by externally-imposed evaluations of performance, such as standardised tests.

While the benefits to members of the group (*e.g.*, a professional association of educators) can be large (*e.g.*, the absence of standardised testing programmes) the costs (*e.g.*, lowered student achievement, a less efficient education or employment system) tend to be diffused over society at large and may not even be noticed by those who bear them. Special interests accrete more and more private benefits (and political power) over time, however, until they become "vested" interests—wealthy, powerful, and entrenched.

Olson's theory is particularly applicable to education in the United States, because its governance is so widely dispersed. Each of the 50 states is

constitutionally responsible for public education and, in 49 states, some governance and taxing authority is further deferred to local school districts, which are typically governed separately from other local units of government. Some national associations of educators maintain substantial memberships in each and every local school district, state legislative district, U.S., congressional district, and television, radio, and newspaper media market. They can saturate the country with the policy-related information they prefer and block out the information dissemination efforts of less powerful individuals or groups that offer contrary points of view.

In the United States, society's understanding of standardised testing may be shrinking. The technical psychometric research literature would seem to be safe. But, the research literature related to testing *policy* (i.e., its administration, programme structure, use, extent, effects, cost, benefits, public opinion, research dissemination) is diminishing. There are simply too few who cite the research literature in any substantial depth or breadth, and too many willing to declare it barren.

The most common debating tactic of testing opponents is to avoid debate (Phelps, 2007a). Whereas scientists seek the scrutiny of their peers in order to confirm (or deny) the value of their work, *advocates* tend to avoid scrutiny, especially when selling falsehoods. Scientists do not circumvent the research literature, but engage it. They respond to rival hypotheses with counterevidence. They confront conflicting scientific results. Advocates, however, simply ignore them. The easiest way to win a debate is by not inviting an opponent. Testing critics rightly fear an open, fair scientific contest.

Indeed, it has become quite common for testing opponents to declare nonexistent an enormous research literature that contradicts their claims. With the help of the fourth estate, they have been fairly successful in eradicating from the collective memory thousands of studies conducted by earnest researchers over the course of a century. In one effort of mine—accumulating studies on the effects of standardised testing—we started out thinking that there were a dozen or so. A few years ago we knew that there were hundreds. Now I know that their number exceeds a thousand.

In the end, however, it will not matter for society's sake if we find ten thousand studies. There will remain other education researchers, prominent and with hugely abundant resources at their disposal—researchers whose work is frequently covered by U.S., education journalists—who will continue to insist that *no* such studies ever existed. It is U.S., education research's dirty big secret: research that generates results that are unpopular among the vested interests can be successfully—and easily—censored and suppressed.

Wildlife conservationists tell us that a biological species cannot survive when mating individuals cannot find each other. When numbers decline to such an extent that predators (or hunters) can more easily find members of the species than can potential mates, the species crosses a demographic threshold and heads towards its inevitable extinction. Those who work with endangered species call

this the “extinction vortex.” Similarly, the censorship and suppression of the research literature on the effects of educational achievement testing has become so successful that it has become difficult to find its progenitors. For example, I may have spent more time than anyone combing the research literature. Nonetheless, I was a few years into my effort before I discovered the work of Frank Dempster (1991, 1997), one of the world’s foremost authorities, or that of Jim Haynie who works in career and technical education. Why did it take me so long to find their work? Their work is not popular among the vested interests in education—they find the benefits of testing to be strong and persistent—thus it is not widely advertised.

One Hundred Years of Research and Experience Left Behind

Indeed, the No Child Left Behind (NCLB) Act, passed by the U.S., Congress in 2002, could have been informed by a cornucopia of research and experience. Instead, it was informed by virtually none. Prior research and experience would have told policymakers that most of the motivational benefits of standardised tests required consequences for the students and not just for the schools. Those stakes needn’t be very high to be effective, but there must be some. As NCLB imposes stakes on schools, but not on students, who knows if the students even try to perform well.

Prior research and experience would have informed policymakers that educators are intelligent people who respond to incentives, and who will game a system if they are given an opportunity to do so. The NCLB Act left many aspects of the test administration process that profoundly affect scores (*e.g.*, incentives and motivation, cut scores, degree of curricular alignment) up for grabs and open to manipulation by local and state officials.

Prior research and experience would have informed policymakers that different tests get different results and one should not expect average scores from different tests to rise and fall in unison over time. Prior research and experience would have informed policymakers that the public was *not* in favour of punishing poorly-performing schools (as NCLB does), but *was* in favour of applying consequences to poorly-performing students and teachers (which NCLB does not).

What are the effects of test-based accountability? The forthcoming *Correcting fallacies about educational and psychological testing* (Phelps, 2008) lists just a small sample of useful, insightful, relevant studies that effectively answered this question, could have informed the design of NCLB, and have been declared by prominent educators to not exist.

Had the policymakers and planners involved in designing the NCLB Act simply read the freely-available research literature instead of funding expensive new studies and waiting for their few results, they would have received more value for their money, gotten more and better information, and gotten it earlier when they actually needed it.

With the single exception of the federal mandate, there was no aspect of the NCLB accountability initiative that had not been tried and studied before. Every

one of the NCLB Act's failings was perfectly predictable, based on decades of prior experience and research. Moreover, there were better alternatives for every characteristic of the programme that had also been tried and studied thoroughly by researchers in psychology, education, and programme evaluation. Yet, policymakers were made aware of none of them.

The resulting scantily-informed public policy includes a national testing programme that would hardly be recognisable anywhere outside of North America. The standardised testing component of NCLB includes no consequences for the students. This sends the subliminal message to the students that they need not work very hard and the testing's largest potential benefit—motivation—is not even accrued.

By contrast, schools are held accountable for students' test performance; they are held responsible for the behaviour of other human beings over whom they have little control. Moreover, the most important potential supporters of testing programmes—classroom teachers and school administrators—are alienated, put into the demeaning position of cajoling students to cooperate.

RELATIONSHIP TO HIGH-STAKES TESTING

Many criterion-referenced tests are also high-stakes tests, where the results of the test have important implications for the individual examinee. Examples of this include high school graduation examinations and licensure testing where the test must be passed to work in a profession, such as to become a physician or attorney. However, being a high-stakes test is not specifically a feature of a criterion-referenced test. It is instead a feature of how an educational or government agency chooses to use the results of the test.

Examples

- Driving tests are criterion-referenced tests, because their goal is to see whether the test taker is skilled enough to be granted a driver's license, not to see whether one test taker is more skilled than another test taker.
- Citizenship tests are usually criterion-referenced tests, because their goal is to see whether the test taker is sufficiently familiar with the new country's history and government, not to see whether one test taker is more knowledgeable than another test taker.

IPSATIVE

Ipsative is a descriptor used in psychology to indicate a specific type of measure in which respondents compare two or more desirable options and pick the one that is most preferred (sometimes called a "forced choice" scale). This is contrasted with measures that use Likert-type scales, in which respondents choose the score (*e.g.*, 1 to 5) which best represents the degree to which they agree with a given statement. "Ipsative Comparisons" are also sometimes used in standardised testing to compare significant differences in subtest scores.

Psychology

While mean scores from Likert-type scales can be compared across individuals, scores from an ipsative measure cannot. To explain, if an individual was equally Extroverted and Conscientious and was assessed on a Likert-type scale, each trait would be evaluated singularly, *i.e.*, a respondent would see the item “I enjoy parties.” and agree or disagree with it to whatever degree reflected his/her preferences.

If the same traits were evaluated on an ipsative measure, respondents would be forced to choose between the two, *i.e.*, a respondent would see the item “Which of these do you agree with more strongly? a) I like parties. b) I keep my workspace neat and tidy.” Ipsative measures may be more useful for evaluating traits within an individual, whereas Likert-type scales are more useful for evaluating traits across individuals. Additionally, ipsative measures may be useful in identifying faking.

Education

In education, ipsative assessment is the practice of assessing present performance against the prior performance of the person being assessed. One place where this might be implemented is in reference to tests used with K-12 students in the United States, where value-added modelling of teacher performance is currently popular.

Ipsative assessment can be contrasted with criterion-referenced assessment and norm-referenced assessment. Ipsative assessment is used in everyday life, and features heavily in physical education and also in computer games. Encouraging pupils to beat their previous scores can take peer pressure out of situations and eliminates the competitive element associated with norm-based referencing. It can be particularly useful for children with learning disabilities and can improve motivation.

CONCEPT INVENTORY

A concept inventory is a criterion-referenced test designed to evaluate whether a student has an accurate working knowledge of a specific set of concepts. To ensure interpretability, it is common to have multiple items that address a single idea.

Typically, concept inventories are organised as multiple-choice tests in order to ensure that they are scored in a reproducible manner, a feature that also facilitates administration in large classes. Unlike a typical, teacher-made multiple-choice test, questions and response choices on concept inventories are the subject of extensive research.

The aims of the research include ascertaining (a) the range of what individuals think a particular question is asking and (b) the most common responses to the questions. Concept inventories are evaluated to ensure test reliability and validity. In its final form, each question includes one correct answer and several distractors. The distractors are incorrect answers that are usually (but not always)

based on students' commonly held misconceptions. Ideally, a score on a criterion-referenced test reflects the amount of content knowledge a student has mastered. Criterion-referenced tests differ from norm-referenced tests in that (in theory) the former is not used to compare an individual's score to the scores of the group. Ordinarily, the purpose of a criterion-referenced test is to ascertain whether a student mastered a predetermined amount of content knowledge; upon obtaining a test score that is at or above a cutoff score, the student can move on to study a body of content knowledge that follows next in a learning sequence. In general, item difficulty values ranging between 30 per cent and 70 per cent are best able to provide information about student understanding. Distractors are often based on ideas commonly held by students, as determined by years of research on misconceptions. Test developers often research student misconceptions by examining students' responses to open-ended essay questions and conducting "think-aloud" interviews with students. The distractors chosen by students help researchers understand student thinking and give instructors insights into students' prior knowledge (and, sometimes, firmly held beliefs). This foundation in research underlies instrument construction and design, and plays a role in helping educators obtain clues about students' ideas, scientific misconceptions, and didaskalogenic, that is, teacher-induced confusions and conceptual lacunae that interfere with learning.

Concept Inventories in Use

The first concept inventory was developed in 1985. It covers the understanding of basic concepts in classical mechanics. Hestenes, Halloun, Wells, and Swackhamer developed the first of the concept inventories to be widely disseminated, the Force Concept Inventory (FCI). The FCI was designed to assess student understanding of the Newtonian concepts of force.

Hestenes (1998) found that while "nearly 80 per cent of the [students completing introductory college physics courses] could state Newton's Third Law at the beginning of the course. FCI data showed that less than 15 per cent of them fully understood it at the end". These results have been replicated in a number of studies involving students at a range of institutions, and have led to greater recognition in the physics education research community of the importance of students' "active engagement" with the materials to be mastered.

Since the development of the FCI, other physics instruments have been developed. These include the Force and Motion Conceptual Evaluation developed by Thornton and Sokoloff and the Brief Electricity and Magnetism Assessment developed by Ding *et al.* In addition to physics, concept inventories have been developed in statistics, chemistry, astronomy, basic biology, natural selection, genetics, engineering, and geoscience.

In many areas, foundational scientific concepts transcend disciplinary boundaries. An example of an inventory that assesses knowledge of such concepts is an instrument developed by Odom and Barrow (1995) to evaluate understanding of diffusion and osmosis. In addition, there are non-multiple

choice conceptual instruments, such as the essay-based approach suggested by Wright *et al.* (1998) and the essay and oral exams used by Nehm and Schonfeld (2008).

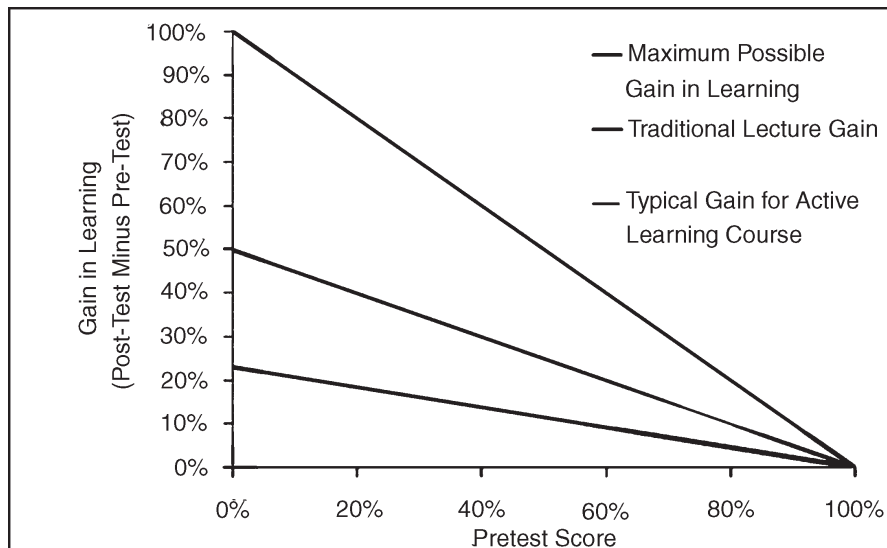


Fig. This is a Schematic of the Hake Plot.

The data collected with a CI are only useful for measuring student learning when the CI is itself valid and reliable. All users are cautioned to carefully review papers for measures of validity and reliability before employing any CI to measure student learning.

Caveats Associated with Concept Inventory use

Some concept inventories are problematic. Some inventories created by scientists do not align with best practices in scale development. Concept inventories created to simply diagnose student ideas may not be viable as research-quality measures of conceptual understanding. Users should be careful to ensure that concept inventories are actually testing conceptual understanding, rather than test-taking ability, language skills, or other abilities that can influence test performance.

The use of multiple-choice exams as concept inventories is not without controversy. The very structure of multiple-choice type concept inventories raises questions involving the extent to which complex, and often nuanced situations and ideas must be simplified or clarified to produce unambiguous responses. For example, a multiple-choice exam designed to assess knowledge of key concepts in natural selection does not meet a number of standards of quality control.

One problem with the exam is that the two members of each of several pairs of parallel items, with each pair designed to measure exactly one key concept in natural selection, sometimes have very different levels of difficulty. Another

problem is that the multiple-choice exam overestimates knowledge of natural selection as reflected in student performance on a diagnostic essay exam and a diagnostic oral exam, two instruments with reasonably good construct validity.

Although scoring concept inventories in the form of essay or oral exams is labour intensive, costly, and difficult to implement with large numbers of students, such exams can offer a more realistic appraisal of the actual levels of students' conceptual mastery as well as their misconceptions. Recently, however, computer technology has been developed that can score essay responses on concept inventories in biology and other domains (Nehm, Ha, and Mayfield, 2011), promising to facilitate the scoring of concept inventories organised as (transcribed) oral exams as well as essays.

DIFFERENCE BETWEEN NORM-REFERENCED AND CRITERION-REFERENCE TESTING

Key Difference: Norm-referenced is a type of test that assesses the test taker's ability and performance against other test takers. Criterion-Reference is a type of test that assesses the test taker's ability to understand a set curriculum.

Norm-Referenced and Criterion-Referenced testing are two of many different types of testing methods that are employed to assess skills of a person. These tests are used to measure performance, but they are relative to different criteria. The scores are also reported in different formats as well as interpreted differently.

Norm-referenced is a type of test that assesses the test taker's ability and performance against other test takers. It could also include a group of test takers against another group of test takers. This is done to differentiate high and low achievers.

The test's content covers a broad area of topics that the test takers are expected to know and the difficulty of the content varies. This test must also be administered in a standardised format. Norm-referenced test helps determine the position of the test taker in a predefined population. Examples of norm-referenced tests include SATs, ACTs, *etc.* These tests do not have a pre-determined curriculum and the topics on the test vary depending on the panel that sets the test.

Criterion-Reference is a type of test that assesses the test taker's ability to understand a set curriculum. In this test, a curriculum is set in the beginning of the class, which is then explained by the instructor. At the end of the lesson, the test is used to determine how much did the test taker understand. This test is commonly used to measure the level of understanding of a test taker before and after an instruction is given.

It can also be used to determine how good the instructor is at teaching the students. The test must have material that is covered in the class by the instructor. The teacher or the instructor sets the test according to the curriculum that was presented. Examples of Criterion-Reference tests include the tests that are given in schools and colleges in classes by a teacher. This helps the teacher determine if the student should pass the class.

	Norm-Referenced	Criterion-Reference
Definition	Norm-Referenced tests measure the performance of one group of test takers against another group of test takers.	Criterion-Reference tests measure the performance of test takers against the criteria covered in the curriculum.
Purpose	To measure how much a test taker knows compared to another student.	To measure how much the test taker known before and after the instruction is finished.
Content	Norm-Referenced tests measure broad skill areas taken from a variety of textbooks and syllabi.	Criterion-Reference tests measure the skills the test taker has acquired on finishing a curriculum.
Item characteristics	Each skill is tested by less than four items. The items vary in difficulty.	Each skill is tested by at least four items to obtain an adequate sample of the student.
Administration	Norm-Referenced tests must be administered in a standardized format.	Criterion-Reference tests need not be administered in a standardized format.
Score reporting	Norm-Referenced test scores are reported in a percentile rank.	Criterion-Reference test scores are reported in categories or percentage.
Score interpretation	In Norm-Referenced tests, if a test taker ranks 95%, it implies that he/she has performed better than 95% of the other test takers.	In Criterion-Reference, the score determines how much of the curriculum is understood by the test taker.

A COMPARISON OF NORM-REFERENCING AND CRITERION-REFERENCING METHODS

The essential characteristic of *norm-referencing* is that students are awarded their grades on the basis of their ranking within a particular cohort. Norm-referencing involves fitting a ranked list of students' 'raw scores' to a pre-determined distribution for awarding grades. Usually, grades are spread to fit a 'bell curve' (a 'normal distribution' in statistical terminology), either by qualitative, informal rough-reckoning or by statistical techniques of varying complexity. For large student cohorts (such as in senior secondary education), statistical moderation processes are used to adjust or standardise student scores to fit a normal distribution. This adjustment is necessary when comparability of scores across different subjects is required (such as when subject scores are added to create an aggregate ENTER score for making university selection decisions).

Norm-referencing is based on the assumption that a roughly similar range of human performance can be expected for any student group. There is a strong culture of norm-referencing in higher education. It is evident in many commonplace practices, such as the expectation that the mean of a cohort's

results should be a fixed percentage year-in year-out (often this occurs when comparability across subjects is needed for the award of prizes, for instance), or the policy of awarding first class honours sparingly to a set number of students, and so on.

In contrast, *criterion-referencing*, as the name implies, involves determining a student's grade by comparing his or her achievements with clearly stated criteria for learning outcomes and clearly stated standards for particular levels of performance. Unlike norm-referencing, there is no pre-determined grade distribution to be generated and a student's grades is in no way influenced by the performance of others.

Theoretically, all students within a particular cohort could receive very high (or very low) grades depending solely on the levels of individuals' performances against the established criteria and standards. The goal of criterion-referencing is to report student achievement against objective reference points that are independent of the cohort being assessed. Criterion-referencing can lead to simple pass-fail grading schema, such as in determining fitness-to-practice in professional fields. Criterion-referencing can also lead to reporting student achievement or progress on a series of key criteria rather than as a single grade or percentage.

Which of these methods is preferable? Mostly, students' grades in universities are decided on a mix of both methods, even though there may not be an explicit policy to do so. In fact, the two methods are somewhat interdependent, more so than the brief explanations above might suggest. Logically, norm-referencing must rely on some initial criterion-referencing, since students' 'raw' scores must presumably be determined in the first instance by assessors who have some objective criteria in mind.

Criterion-referencing, on the other hand, appears more educationally defensible. But criterion-referencing may be very difficult, if not impossible, to implement in a pure form in many disciplines. It is not always possible to be entirely objective and to comprehensively articulate criteria for learning outcomes: some subjectivity in setting and interpreting levels of achievement is inevitable in higher education. This being the case, sometimes the best we can hope for is to compare individuals' achievements relative to their peers.

Norm-referencing, on its own — and if strictly and narrowly implemented — is undoubtedly unfair. With norm-referencing, a student's grade depends — to some extent at least — not only on his or her level of achievement, but also on the achievement of other students. This might lead to obvious inequities if applied without thought to any other considerations. For example, a student who fails in one year may well have passed in other years! The potential for unfairness of this kind is most likely in smaller student cohorts, where norm-referencing may force a spread of grades and exaggerate differences in achievement. Alternatively, norm-referencing might artificially compress the range of difference that actually exists.

Criterion-referencing is worth aspiring towards. Criterion-referencing requires giving thought to expected learning outcomes: it is transparent for

students, and the grades derived should be defensible in reasonably objective terms – students should be able to trace their grades to the specifics of their performance on set tasks. Criterion-referencing lays an important framework for student engagement with the learning process and its outcomes.

Recognising, however, that some degree of subjectivity is inevitable in higher education, it is also worthwhile to monitor grade distributions – in other words, to use a modest process of norm-referencing to watch the outcomes of a predominantly criterion-referenced grading model. In doing so, if it is believed too many students are receiving low grades, or too many students are receiving high grades, or the distribution is in some way oddly spread, then this might suggest something is amiss and the assessment process needs looking at. There may be, for instance, a problem with the overall degree of difficulty of the assessment tasks (for example, not enough challenging examination questions, or too few, or assignment tasks that fail to discriminate between students with differing levels of knowledge and skills). There might also be inconsistencies in the way different assessors are judging student work. Best practice in grading in higher education involves striking a balance between criterion-referencing and norm-referencing. This balance should be strongly oriented towards criterion-referencing as the primary and dominant principle.

PHYSICAL FITNESS TESTS

A physical fitness test is a test designed to measure physical strength, agility, and endurance. They are commonly employed in educational institutions as part of the physical education curriculum, in medicine as part of diagnostic testing, and as eligibility requirements in fields that focus on physical ability such as military or police. Throughout the 20th century, scientific evidence emerged demonstrating the usefulness of strength training and aerobic exercise in maintaining overall health, and more agencies began to incorporate standardised fitness testing. In the United States, the President's Council on Youth Fitness was established in 1956 as a way to encourage and monitor fitness in schoolchildren. Common tests include timed running or the multi-stage fitness test, and numbers of push-ups, sit-ups/abdominal crunches and pull-ups that the individual can perform. More specialised tests may be used to test ability to perform a particular job or role.

Performance Tests

A performance test is an assessment that requires an examinee to actually perform a task or activity, rather than simply answering questions referring to specific parts. The purpose is to ensure greater fidelity to what is being tested. An example is a behind-the-wheel driving test to obtain a driver's license. Rather than only answering simple multiple-choice items regarding the driving of an automobile, a student is required to actually drive one while being evaluated.

Performance tests are commonly used in workplace and professional applications, such as professional certification and licensure. When used for

personnel selection, the tests might be referred to as a work sample. A licensure example would be cosmetologists being required to demonstrate a haircut or manicure on a live person. The Group-Bourdon test is one of a number of psychometric tests which trainee train drivers in the UK are required to pass.

Some performance tests are simulations. For instance, the assessment to become certified as an ophthalmic technician includes two components, a multiple-choice examination and a computerised skill simulation. The examinee must demonstrate the ability to complete seven tasks commonly performed on the job, such as retinoscopy, that are simulated on a computer.

Test Preparations

From the perspective of a test developer, there is great variability with respect to time and effort needed to prepare a test. Likewise, from the perspective of a test taker, there is also great variability with respect to the time and needed to obtain a desired grade or score on any given test. When a test developer constructs a test, the amount of time and effort is dependent upon the significance of the test itself, the proficiency of the test taker, the format of the test, class size, deadline of test, and experience of the test developer.

The process of test construction has been greatly aided in several ways. For one, many test developers were themselves students at one time, and therefore are able to modify or outright adopt test questions from their previous tests. In some countries such as the United States, book publishers often provide teaching packages that include test banks to university instructors who adopt their published books for their courses. These test banks may contain up to four thousand sample test questions that have been peer-reviewed and time tested. The instructor who chooses to use this testbank would only have to select a fixed number of test questions from this test bank to construct a test.

As with test constructions, the time needed for a test taker to prepare for a test is dependent upon the frequency of the test, the test developer, and the significance of the test. In general, nonstandardised tests that are short, frequent, and do not constitute a major portion of the test taker's overall course grade or score require do not require the test taker to spend great amounts preparing for the test. Conversely, nonstandardised tests that are long, infrequent, and do constitute a major portion of the test taker's overall course grade or score usually require the test taker to spend great amounts preparing for the test.

To prepare for a nonstandardised test, test takers may rely upon their reference books, class or lecture notes, Internet, and past experience to prepare for the test. Test takers may also use various learning aids to study for tests such as flash cards and mnemonics. Test takers may even hire tutors to coach them through the process so that they may increase the probability of obtaining a desired test grade or score. Finally, test takers may rely upon past copies of a test from previous years or semesters to study for a future test. These past tests may be provided by a friend or a group that has copies of previous tests or from instructors and their institutions.

Unlike nonstandardised test, the time needed by test takers to prepare for standardised tests are less variable and usually considerable. This is because standardised tests are usually uniformed in scope, format, and difficulty and often have important consequences with respect to a test taker's future such as a test taker's eligibility to attend a specific university programme or to enter a desired profession. It is not unusual for test takers to prepare for standardised tests by relying upon commercially available books that provide in-depth coverage of the standardised test or compilations of previous tests (*e.g.*, 10 year series in Singapore). In many countries, test takers even enroll in test preparation centres or cram schools that provide extensive or supplementary instructions to test takers to help them better prepare for a standardised test. Finally, in some countries, instructors and their institutions have also played a significant role in preparing test takers for a standardised test.

Cheating on Tests

Cheating on a test is the process of using unauthorised means or methods for the purpose of obtaining a desired test score or grade. This may range from bringing and using notes during a closed book examination, copying another test taker's answer or choice of answers during an individual test, or even sending a paid proxy to take the test.

Several common methods have been employed to combat cheating. They include the use of multiple proctors or invigilators during a testing period to monitor test takers. Test developers may construct multiple variants of the same test to be administered to different test takers at the same time. In some cases, instructors themselves may not administer their own tests but will leave the task to other instructors or invigilators, which may mean that the invigilators do not know the candidates, and thus some form of identification may be required. Finally, instructors or test providers may compare the answers of suspected cheaters on the test themselves to determine if cheating did occur.

Support and Criticisms of Tests

Despite their widespread use, the validity, quality, or use of tests, particularly standardised tests in education have continued to be widely supported or criticized. Like the tests themselves, supports and criticisms of tests are often varied and may come from a variety of sources such as parents, test takers, instructors, business groups, universities, or governmental watchdogs.

Supporters of standardised tests in education often provide the following reasons for promoting testing in education:

- Feedback or diagnosis of test taker's performance
- Fair and efficient
- Promotes accountability
- Prediction and selection
- Improves performance

Critics of standardised tests in education often provide the following reasons for revising or removing standardised tests in education:

- Narrows curricular format and encourages teaching to the test.
- Poor predictive quality.
- Grade inflation of test scores or grades.
- Culturally or socioeconomically biased.

Aptitude

An aptitude is a component of a competency to do a certain kind of work at a certain level, which can also be considered “talent”. Aptitudes may be physical or mental. Aptitude is not knowledge, understanding, learned or acquired abilities (skills) or attitude. The innate nature of aptitude is in contrast to achievement, which represents knowledge or ability that is gained.

Intelligence

Aptitude and intelligence quotient are related, and in some ways opposite views of human mental ability. Whereas intelligence quotient sees intelligence as being a single measurable characteristic affecting all mental ability, aptitude refers to one of many different characteristics which can be independent of each other, such as aptitude for military flight, air traffic control, or computer programming. This is more similar to the theory of multiple intelligences.

Concerning a single measurable characteristic affecting all mental ability, analysis of any group of intelligence test scores will nearly always show them to be highly correlated. The U.S., Department of Labour’s General Learning Ability, for instance, is determined by combining Verbal, Numerical and Spatial aptitude subtests. In a given person some are low and others high. In the context of an aptitude test the “high” and “low” scores are usually not far apart, because all ability test scores tend to be correlated.

Aptitude is better applied intra-individually to determine what tasks a given individual is more skilled at performing. Inter-individual aptitude differences are typically not very significant due to IQ differences.

Of course this assumes individuals have not already been pre-screened for aptitude through some other process such as SAT scores, GRE scores, or finishing medical school.

Combined Aptitude and Knowledge Tests

Tests that assess learned skills or knowledge are frequently called achievement tests. However, certain tests can assess both types of constructs. An example that leans both ways is the Armed Services Vocational Aptitude Battery (ASVAB), which is given to recruits entering the armed forces of the United States. Another is the SAT, which is designed as a test of aptitude for college in the United States, but has achievement elements. For example, it tests mathematical reasoning, which depends both on innate mathematical ability and education received in mathematics.

Aptitude tests can typically be grouped according to the type of cognitive ability they measure:

1. *Fluid intelligence*: the ability to think and reason abstractly, effectively solve problems and think strategically. It's more commonly known as 'street smarts' or the ability to 'quickly think on your feet'. Examples of what employers can learn from your fluid intelligence about your suitability for the role for which you are applying
2. *Crystallised intelligence*: the ability to learn from past experiences and relevant learning, and to apply this learning to work-related situation. Work situations that require crystallised intelligence include producing and analysing written reports, comprehending work instructions, using numbers as a tool to make effective decisions, *etc.*

6

Evaluation in a Democratic School

Sudbury model of democratic education schools do not perform and do not offer evaluations, assessments, transcripts, or recommendations, asserting that they do not rate people, and that school is not a judge; comparing students to each other, or to some standard that has been set is for them a violation of the student's right to privacy and to self-determination. Students decide for themselves how to measure their progress as self-starting learners as a process of self-evaluation: real lifelong learning and the proper educational evaluation for the 21st Century, they adduce.

According to Sudbury schools(Riaz Institute of education and research.)..., this policy does not cause harm to their students as they move on to life outside the school. However, they admit it makes the process more difficult, but that such hardship is part of the students learning to make their own way, set their own standards and meet their own goals.

The no-grading and no-rating policy helps to create an atmosphere free of competition among students or battles for adult approval, and encourages a positive co-operative environment amongst the student body.

The final stage of a Sudbury education, should the student choose to take it, is the graduation thesis. Each student writes on the topic of how they have prepared themselves for adulthood and entering the community at large. This thesis is submitted to the Assembly, who reviews it.

The final stage of the thesis process is an oral defence given by the student in which they open the floor for questions, challenges and comments from all Assembly members. At the end, the Assembly votes by secret ballot on whether or not to award a diploma.

CONSEQUENCES OF LEVEL OF MEASUREMENT

Why are we so interested in the type of scale that measures a dependent variable? The crux of the matter is the relationship between the variable's level of measurement and the statistics that can be meaningfully computed with that variable. For example, consider a hypothetical study in which 5 children are asked to choose their favourite colour from blue, red, yellow, green, and purple. The researcher codes the results as follows:

COLOUR CODE

Blue	1
Red	2
Yellow	3
Green	4
Purple	5

This means that if a child said her favourite colour was "Red," then the choice was coded as "2," if the child said her favourite colour was "Purple," then the response was coded as 5, and so forth. Consider the following hypothetical data:

Subject Colour Code

1. Blue 1
2. Blue 1
3. Green 4
4. Green 4
5. Purple 5

Each code is a number, so nothing prevents us from computing the average code assigned to the children. The average happens to be 3, but you can see that it would be senseless to conclude that the average favourite colour is yellow (the colour with a code of 3).

Such non-sense arises because favourite colour is a nominal scale, and taking the average of its numerical labels is like counting the number of letters in the name of a snake to see how long the beast is.

Does it make sense to compute the mean of numbers measured on an ordinal scale? This is a difficult question, one that statisticians have debated for decades. You will be able to explore this issue yourself in a simulation shown in the next section and reach your own conclusion.

The prevailing (but by no means unanimous) opinion of statisticians is that for almost all practical situations, the mean of an ordinally-measured variable is a meaningful statistic. However, as you will see in the simulation, there are extreme situations in which computing the mean of an ordinally-measured variable can be very misleading.

LEVEL OF MEASUREMENT IS USED FOR PSYCHOLOGICAL VARIABLES

Rating scales are used frequently in psychological research. For example, experimental subjects may be asked to rate their level of pain, how much they like a consumer product, their attitudes about capital punishment, their confidence in an answer to a test question. Typically these ratings are made on a 5-point or a 7-point scale. These scales are ordinal scales since there is no assurance that a given difference represents the same thing across the range of the scale. For example, there is no way to be sure that a treatment that reduces pain from a rated pain level of 3 to a rated pain level of 2 represents the same level of relief as a treatment that reduces pain from a rated pain level of 7 to a rated pain level of 6.

In memory experiments, the dependent variable is often the number of items correctly recalled. What scale of measurement is this? You could reasonably argue that it is a ratio scale. First, there is a true zero point: some subjects may get no items correct at all. Moreover, a difference of one represents a difference of one item recalled across the entire scale. It is certainly valid to say that someone who recalled 12 items recalled twice as many items as someone who recalled only 6 items.

But number-of-items recalled is a more complicated case than it appears at first. Consider the following example in which subjects are asked to remember as many items as possible from a list of 10. Assume that (a) there are 5 easy items and 5 difficult items, (b) half of the subjects are able to recall all the easy items and different numbers of difficult items, while (c) the other half of the subjects are unable to recall any of the difficult items but they do remember different numbers of easy items. Some sample data are shown below.

Subject Easy Items Difficult Items Score

A	0	0	1	1	0	0	0	0	0	0	2
B	1	0	1	1	0	0	0	0	0	0	3
C	1	1	1	1	1	1	1	0	0	0	7
D	1	1	1	1	1	0	1	1	0	1	8

Let's compare:

1. The difference between Subject A's score of 2 and Subject B's score of 3 with
2. The difference between Subject C's score of 7 and Subject D's score of 8.

The former difference is a difference of one easy item; the latter difference is a difference of one difficult item. Do these two differences necessarily signify the same difference in memory? We are inclined to respond "No" to this question since only a little more memory may be needed to retain the additional easy item whereas a lot more memory may be needed to retain the additional hard item. The general point is that it is often inappropriate to consider psychological measurement scales as either interval or ratio.

EDUCATION AS DEVELOPMENT

THE CONDITIONS OF GROWTH

In directing the activities of the young, society determines its own future in determining that of the young. Since the young at a given time will at some later date compose the society of that period, the latter's nature will largely turn upon the direction children's activities were given at an earlier period. This cumulative movement of action towards a later result is what is meant by growth.

The primary condition of growth is immaturity. This may seem to be a mere truism — saying that a being can develop only in some point in which he is undeveloped. But the prefix "im" of the word immaturity means something positive, not a mere void or lack. It is noteworthy that the terms "capacity" and "potentiality" have a double meaning, one sense being negative, the other positive.

Capacity may denote mere receptivity, like the capacity of a quart measure. We may mean by potentiality a merely dormant or quiescent state — a capacity to become something different under external influences. But we also mean by capacity an ability, a power; and by potentiality potency, force. Now when we say that immaturity means the possibility of growth, we are not referring to absence of powers which may exist at a later time; we express a force positively present — the ability to develop.

Our tendency to take immaturity as mere lack, and growth as something which fills up the gap between the immature and the mature is due to regarding childhood comparatively, instead of intrinsically. We treat it simply as a privation because we are measuring it by adulthood as a fixed standard. This fixes attention upon what the child has not, and will not have till he becomes a man. This comparative standpoint is legitimate enough for some purposes, but if we make it final, the question arises whether we are not guilty of an overweening presumption. Children, if they could express themselves articulately and sincerely, would tell a different tale; and there is excellent adult authority for the conviction that for certain moral and intellectual purposes adults must become as little children. The seriousness of the assumption of the negative quality of the possibilities of immaturity is apparent when we reflect that it sets up as an ideal and standard a static end. The fulfillment of growing is taken to mean an accomplished growth: that is to say, an Ungrowth, something which is no longer growing. The futility of the assumption is seen in the fact that every adult resents the imputation of having no further possibilities of growth; and so far as he finds that they are closed to him mourns the fact as evidence of loss, instead of falling back on the achieved as adequate manifestation of power. Why an unequal measure for child and man?

Taken absolutely, instead of comparatively, immaturity designates a positive force or ability, — the power to grow. We do not have to draw out or educe positive activities from a child, as some educational doctrines would have it. Where there is life, there are already eager and impassioned activities. Growth

is not something done to them; it is something they do. The positive and constructive aspect of possibility gives the key to understanding the two chief traits of immaturity, dependence and plasticity.

It sounds absurd to hear dependence spoken of as something positive, still more absurd as a power. Yet if helplessness were all there were in dependence, no development could ever take place. A merely impotent being has to be carried, forever, by others. The fact that dependence is accompanied by growth in ability, not by an ever increasing lapse into parasitism, suggests that it is already something constructive. Being merely sheltered by others would not promote growth.

It would only build a wall around impotence. With reference to the physical world, the child is helpless. He lacks at birth and for a long time thereafter power to make his way physically, to make his own living. If he had to do that by himself, he would hardly survive an hour. On this side his helplessness is almost complete. The young of the brutes are immeasurably his superiors. He is physically weak and not able to turn the strength which he possesses to coping with the physical environment. The thoroughgoing character of this helplessness suggests, however, some compensating power. The relative ability of the young of brute animals to adapt themselves fairly well to physical conditions from an early period suggests the fact that their life is not intimately bound up with the life of those about them. They are compelled, so to speak, to have physical gifts because they are lacking in social gifts. Human infants, on the other hand, can get along with physical incapacity just because of their social capacity. We sometimes talk and think as if they simply happened to be physically in a social environment; as if social forces exclusively existed in the adults who take care of them, they being passive recipients.

If it were said that children are themselves marvelously endowed with power to enlist the cooperative attention of others, this would be thought to be a backhanded way of saying that others are marvelously attentive to the needs of children. But observation shows that children are gifted with an equipment of the first order for social intercourse. Few grown-up persons retain all of the flexible and sensitive ability of children to vibrate sympathetically with the attitudes and doings of those about them. Inattention to physical things is accompanied by a corresponding intensification of interest and attention as to the doings of people. The native mechanism of the child and his impulses all tend to facile social responsiveness. The statement that children, before adolescence, are egotistically self-centred, even if it were true, would not contradict the truth of this statement. It would simply indicate that their social responsiveness is employed on their own behalf, not that it does not exist. But the statement is not true as matter of fact. The facts which are cited in support of the alleged pure egoism of children really show the intensity and directness with which they go to their mark.

If the ends which form the mark seem narrow and selfish to adults, it is only because adults have mastered these ends, which have consequently ceased to

interest them. Most of the remainder of children's alleged native egoism is simply an egoism which runs counter to an adult's egoism. To a grown-up person who is too absorbed in his own affairs to take an interest in children's affairs, children doubtless seem unreasonably engrossed in their own affairs.

From a social standpoint, dependence denotes a power rather than a weakness; it involves interdependence. There is always a danger that increased personal independence will decrease the social capacity of an individual. In making him more self-reliant, it may make him more self-sufficient; it may lead to aloofness and indifference. It often makes an individual so insensitive in his relations to others as to develop an illusion of being really able to stand and act alone — an unnamed form of insanity which is responsible for a large part of the remediable suffering of the world.

The specific adaptability of an immature creature for growth constitutes his plasticity. This is something quite different from the plasticity of putty or wax. It is not a capacity to take on change of form in accord with external pressure. It lies near the pliable elasticity by which some persons take on the colour of their surroundings while retaining their own bent. But it is something deeper than this. It is essentially the ability to learn from experience; the power to retain from one experience something which is of avail in coping with the difficulties of a later situation. This means power to modify actions on the basis of the results of prior experiences, the power to develop dispositions. Without it, the acquisition of habits is impossible.

It is a familiar fact that the young of the higher animals, and especially the human young, have to learn to utilize their instinctive reactions. The human being is born with a greater number of instinctive tendencies than other animals. But the instincts of the lower animals perfect themselves for appropriate action at an early period after birth, while most of those of the human infant are of little account just as they stand. An original specialized power of adjustment secures immediate efficiency, but, like a railway ticket, it is good for one route only. A being who, in order to use his eyes, ears, hands, and legs, has to experiment in making varied combinations of their reactions, achieves a control that is flexible and varied. A chick, for example, pecks accurately at a bit of food in a few hours after hatching.

This means that definite coordination of activities of the eyes in seeing and of the body and head in striking are perfected in a few trials. An infant requires about six months to be able to gauge with approximate accuracy the action in reaching which will coordinate with his visual activities; to be able, that is, to tell whether he can reach a seen object and just how to execute the reaching. As a result, the chick is limited by the relative perfection of its original endowment. The infant has the advantage of the multitude of instinctive tentative reactions and of the experiences that accompany them, even though he is at a temporary disadvantage because they cross one another.

In learning an action, instead of having it given ready-made, one of necessity learns to vary its factors, to make varied combinations of them, according to

change of circumstances. A possibility of continuing progress is opened up by the fact that in learning one act, methods are developed good for use in other situations. Still more important is the fact that the human being acquires a habit of learning. He learns to learn.

The importance for human life of the two facts of dependence and variable control has been summed up in the doctrine of the significance of prolonged infancy. ¹ This prolongation is significant from the standpoint of the adult members of the group as well as from that of the young.

The presence of dependent and learning beings is a stimulus to nurture and affection. The need for constant continued care was probably a chief means in transforming temporary cohabitations into permanent unions.

It certainly was a chief influence in forming habits of affectionate and sympathetic watchfulness; that constructive interest in the well-being of others which is essential to associated life. Intellectually, this moral development meant the introduction of many new objects of attention; it stimulated foresight and planning for the future. Thus there is a reciprocal influence. Increasing complexity of social life requires a longer period of infancy in which to acquire the needed powers; this prolongation of dependence means prolongation of plasticity, or power of acquiring variable and novel modes of control. Hence it provides a further push to social progress.

TYPES OF SCALES

Before we can conduct a statistical analysis, we need to measure our dependent variable. Exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favourable," "somewhat favourable," *etc.*). For a dependent variable such as "favourite colour," you can simply note the colour-word (like "red") that the subject offers.

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called "scale types," or just "scales," and are described in this section.

NOMINAL SCALES

When measuring using a nominal scale, one simply names or categorizes responses. Gender, handedness, favourite colour, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favourite colour, there is no sense in which green is placed "ahead of" blue. Responses are merely categorized. Nominal scales embody the lowest level of measurement.

Ordinal Scales

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the dependent variable.

For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens. Such an assertion reflects the first person's use of a verbal label that comes later in the list than the label chosen by the second person.

On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is probably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied."

Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction. Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

What if the researcher had measured satisfaction by asking consumers to indicate their level of satisfaction by choosing a number from one to four? Would the difference between the responses of one and two necessarily reflect the same difference in satisfaction as the difference between the responses two and three? The answer is No. Changing the response format to numbers does not change the meaning of the scale. We still are in no position to assert that the mental step from 1 to 2 (for example) is the same as the mental step from 3 to 4.

Interval Scales

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not

represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label “zero” is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures.

For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the “zero” label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is “twice as hot” as 40 degrees. Such a claim would depend on an arbitrary decision about where to “start” the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

Ratio Scales

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. You can think of a ratio scale as the three earlier scales rolled up in one.

Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

The Fahrenheit scale for temperature has an arbitrary zero point and is therefore not a ratio scale. However, zero on the Kelvin scale is absolute zero. This makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, *etc.*). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money.

Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents (or that Bill Gates has a million times more money than you do).

DIFFERENCE BETWEEN CATEGORICAL, ORDINAL AND INTERVAL VARIABLES

In talking about variables, sometimes you hear variables being described as categorical (or sometimes nominal), or ordinal, or interval. Below we will define these terms and explain why they are important.

CATEGORICAL

A categorical variable (sometimes called a nominal variable) is one that has two or more categories, but there is no intrinsic ordering to the categories. For example, gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories. Hair colour is also a categorical variable having a number of categories (blonde, brown, brunette, red, *etc.*) and again, there is no agreed way to order these from highest to lowest. A purely categorical variable is one that simply allows you to assign categories but you cannot clearly order the variables. If the variable has a clear ordering, then that variable would be an ordinal variable, as described below.

Ordinal

An ordinal variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables. For example, suppose you have a variable, economic status, with three categories (low, medium and high). In addition to being able to classify people into these three categories, you can order the categories as low, medium and high. Now consider a variable like educational experience (with values such as elementary school graduate, high school graduate, some college and college graduate). These also can be ordered as elementary school, high school, some college, and college graduate. Even though we can order these from lowest to highest, the spacing between the values may not be the same across the levels of the variables.

Say we assign scores 1, 2, 3 and 4 to these four levels of educational experience and we compare the difference in education between categories one and two with the difference in educational experience between categories two and three, or the difference between categories three and four. The difference between categories one and two (elementary and high school) is probably much bigger than the difference between categories two and three (high school and some college). In this example, we can order the people in level of educational experience but the size of the difference between categories is inconsistent (because the spacing between categories one and two is bigger than categories two and three). If these categories were equally spaced, then the variable would be an interval variable.

Interval

An interval variable is similar to an ordinal variable, except that the intervals between the values of the interval variable are equally spaced. For example, suppose you have a variable such as annual income that is measured in dollars, and we have three people who make \$10,000, \$15,000 and \$20,000. The second person makes \$5,000 more than the first person and \$5,000 less than the third person, and the size of these intervals is the same. If there were two other people who make \$90,000 and \$95,000, the size of that interval between these two people is also the same (\$5,000).

Matter Whether a Variable is Categorical, Ordinal or Interval

Statistical computations and analyses assume that the variables have a specific levels of measurement. For example, it would not make sense to compute an average hair colour. An average of a categorical variable does not make much sense because there is no intrinsic ordering of the levels of the categories. Moreover, if you tried to compute the average of educational experience as defined in the ordinal section above, you would also obtain a non-sensical result. Because the spacing between the four levels of educational experience is very uneven, the meaning of this average would be very questionable.

In short, an average requires a variable to be interval. Sometimes you have variables that are “in between” ordinal and interval, for example, a five-point likert scale with values “strongly agree”, “agree”, “neutral”, “disagree” and “strongly disagree”. If we cannot be sure that the intervals between each of these five values are the same, then we would not be able to say that this is an interval variable, but we would say that it is an ordinal variable. However, in order to be able to use statistics that assume the variable is interval, we will assume that the intervals are equally spaced.

Matter if My Dependent Variable is Normally Distributed

When you are doing a t-test or ANOVA, the assumption is that the distribution of the sample means are normally distributed. One way to guarantee this is for the distribution of the individual observations from the sample to be normal. However, even if the distribution of the individual observations is not normal, the distribution of the sample means will be normally distributed if your sample size is about 30 or larger.

EVALUATION IN EDUCATION

Evaluation in education is a multifaceted process aimed at assessing the effectiveness, quality, and impact of educational programs, policies, and practices. It encompasses various methods and techniques for gathering, analyzing, and interpreting data to inform decision-making and improve educational outcomes. Evaluation plays a crucial role in determining the extent to which educational goals and objectives are being achieved and identifying areas for improvement. It involves assessing student learning and performance, as well as evaluating the efficacy of instructional methods, curriculum materials, and learning environments. Additionally, evaluation in education extends beyond the classroom to encompass broader educational initiatives, such as school reform efforts, educational policies, and systemic interventions. Key components of evaluation in education include setting clear and measurable goals, selecting appropriate evaluation methods and instruments, collecting relevant data, analyzing findings, and using evaluation results to inform decision-making, policy development, and program improvement. Overall, evaluation in education is essential for ensuring accountability, promoting continuous improvement, and enhancing the quality and effectiveness of educational systems and practices. The book on Evaluation in Education provides comprehensive insights and practical strategies for conducting effective evaluations to assess and enhance the quality of educational programs, policies, and practices.



Dr. Sangeeta Agarwal did an M.A. in English, M.Com in Business Management, B.Ed. & M.Ed. from MGSU, Bikaner (Rajasthan). Her interest towards research led her to obtain Ph.D. Degree in Education from Tanta University, Sriganganagar. Currently, she is working as an Associate Professor in Education Department, Tanta University, Sriganganagar (Raj.) She has published many papers in National Journals. She has also participated in National Seminars & presented her papers. She has teaching experience of more than 10 years. Under her guidance 6 students were awarded Ph.D. Degree.



**ACADEMIC
UNIVERSITY PRESS**

4378/4-B, Murarilal Street, Ansari Road, Daryaganj, New Delhi-110002
Phone : +91-11-23281685, 41043100, Fax: +91-11-23270680
E-Mail: academicuniversitypress@gmail.com

ISBN-978-93-6284-007-3

