Chapter 2: Early approaches to quantification of learning and scale development

> It seems that too often that we, and students in particular, are remiss in studying the history of our field. This is unfortunate because this historical background provides the framework within which we can interpret the value of current work as well as allowing us to assess the progress of our science.
>
> De Ayala, 2008, p. 209

The history of the early development of educational assessment is important because it reveals some of the thinking behind early attempts to make teacher judgement assessments consistent. One approach to the quantification of learning sought to identify examples of student work of increasing quality and to allocate values to the examples. These examples and their values were used as references for teachers to judge the quality of other cases of student work. Other approaches to quantifying learning required direct responses from students to test items. In both forms of assessment, the early developers created scales to order and space the quality of the work or the abilities of students. The scaling processes in these early developments indicate that the potential was there to better integrate the role of the teacher as assessor into the teaching/learning cycle than ultimately occurred. Two early innovators, Thorndike and Courtis, feature most significantly in the chapter and interact with a number of the other contributors, many of whom were students of Thorndike.

The chapter considers how information about learning in schools was obtained by the first researchers. Initial student examinations and surveys, and the instruments and processes adopted to carry them out, provide a context for the developments in the first 25 years of the twentieth century. As is the case of other scientific endeavours the earlier works inspired those who came next to refine and develop those ideas. The chapter describes the processes for obtaining information, and insights from early educational surveys and early approaches to creating scales, that led to the quantification of learning.

Placing examples or students on a scale required units for those scales. These units represented the value of the example or the score of the student, thereby providing quantitative values for learning. Thorndike, and later Thurstone, developed approaches to this. Interestingly, in hindsight, their solutions can be shown to bear a linear relationship to the log odds unit, adopted now as the logit of modern Item Response Theory and the Rasch model. In a consideration of teacher judgement assessment, this thesis argues that units are required for teachers to indicate learning status. The early history of quantifying learning provides examples that link the initial steps to improve teacher assessment consistency to the

potentially broader application of teacher judgement assessment as the prime source of learning information about students.

Timeline of the key examples considered

A number of case studies of the 19th and early 20th century illustrate the development of approaches to understanding classroom learning processes that were taken by administrators, teachers or researchers. Initially these take the form of examinations and surveys. New processes evolved from these into standards-referenced examples, drawing their examples from authentic student work. Mixed in with these developments is what are recognised today as pencil and paper tests. Table 2.1 below summarises the key periods and developments tracked in this chapter.

Table 2.1 Timeline of historical developments in assessment considered

Year	Development
1845	Boston: common exam on one day, holistic judgement of writing.
1850-1862	Greenwich: Rev. Fisher employs his scale book reference system for a range of subjects.
1892-1893	US: Rice's first survey of schools; mainly observational.
1897	US: Rice's survey of spelling; development of systematic standard approaches to data collection.
1902	US: Rice's survey of arithmetic.
1908	Stone, influenced by Rice's surveys, develops two arithmetic tests, 'fundamentals' and 'reasoning' to survey arithmetic in a number of school systems.
1909 -1911	Courtis replicates Stone's survey in his own school and then refines and expands the general approach to better understand improvement across grades. Courtis's tests become popular and provide comparison data back to Courtis for reference by teachers in their classroom use of tests.
1910	Thorndike develops the Handwriting scale.
1912	Hillegas (student of Thorndike) develops a scaled (holistic) judgment approach for the quality of prose composition Thorndike argues benefits of scale positioned standards as approaches to measurement in education.
1913	Thorndike adds items to the Hillegas (prose quality) scale.
1914	Ballou develops Harvard-Newton Scales; an alternative to the Hillegas scale. A set of instruments for composition, one for each of the four discourses. Thorndike (with Gray) develops a reading ability scale ('scale a' for visual vocabulary for single words) and a reading comprehension scale ('scale Alpha' for measuring the understanding of sentences). Courtis develops a composition scale similar to that of Rice.

Year	Development
1916	Hudelson advocates a standard running from 10 to 120 in degrees of difficulty; with the minimum for the first grade being 10, second 20, up to 120 for the twelfth year. Appears to be first conception of learning into *levels*. Trabue Completion test: an indicator of the ability to think about words and language. Items are scaled, based on Thorndike's approach and four parallel tests developed. Used as pre- post-tests to gauge annual progress.
1916-1923	Various new scales developed.
1917	Trabue adds items to the Hillegas scale (the Nassau County Supplement).
1925	Thurstone proposes approach to item scaling and visual representation; the first item map.
1950s	Rasch addresses ways of connecting test data over time and develops sample - independent establishment of item difficulty. Leads to new approaches to item scaling, test data analysis and scaling units.
(1984)	Engelhard, using Trabue's 1916 data, establishes that early approaches to scaling and item invariance by Thorndike and Thurstone approximate the Rasch approach.

1845 Massachusetts: the first system wide examination process in the US

Systematic approaches to examinations of students in a standardised form in the US are traced back to Massachusetts and are reported by Mann (Mann, 1845, reprinted in full in Caldwell & Courtis, 1925). While historical references emphasise the importance of Mann (e.g. Butts, 1978; Johnson, Dupuis, Musial, Hall, & Gollnick, 2002), the actual credit for the reported improvement in examination approaches goes to a close confidant of Mann, Samuel G. Howe (Good, 1926), a member of the School Survey Committees of Boston. These Committees, one for Grammar Schools and a second for the Writing Schools, had responsibility for reporting annually on each school, somewhat akin to one of the roles of Her Majesty's Inspectors in 19th Century Britain, although in Boston these were unpaid committee members. Sub-committees developed reports following one-day visits (Caldwell & Courtis, 1925, p. 195).

Prior to 1845 inspections had relied on oral tests and observations and were perfunctorily reported. In 1845 the procedures for the survey committees for reporting on each of the Grammar Schools and Writing Schools were radically modified to include a written examination of students (Caldwell & Courtis, 1925, p. 26).

A survey committee of three was established to report on Grammar Schools. The committee made clear its reasons for the written examination approach. Independence of process and an evidence base were seen as fair in reporting on the schools, though implied in the process was a mistrust of school staff. The committee applied assessment processes to give the same advantages to all (avoiding leading questions) so as

> to ascertain with certainty, what the scholars did not know, as well as what they did know; to test their readiness at expressing their ideas upon paper; to have positive and undeniable evidence of their ability or inability to construct sentences grammatically, to punctuate them, and to spell the words. (Committee Report, 1845 cited in Caldwell & Courtis, 1925, p. 26)

Historically, this represents a key change in assessment practice. The logistics adopted to achieve the fairness objective were comprehensive. Mann (1845) acknowledged parallel developments in written examinations in Europe and Great Britain. The processes for reporting on the quality of schools at that time in Europe appear much less comprehensive than those of the Boston committees. Reports on the processes of Her Majesty's Inspectors of the time suggest a model closer to the pre-revision Boston model (Arnold, 1889; Sneyd-Kynnersley, 1913; Wyatt, 1917). The examination boards of Great Britain, for example, were not established until 1857 for Oxford (Oxford University Archives) and 1858 for Cambridge (Cambridge Assessment Archives).

The Boston School Committees' reports document the beginning elements of school system wide approaches to examinations. Concepts of common test items, external control of timing, the time allowed for the test, security, approaches to analysis and a public report on the results were precursors to general system-wide testing as it has evolved today. The process did not provide assessment or pedagogical support to the classroom but was sophisticated for its time. It sets the scene for the external survey testing approach as a way of understanding the quality of schooling through a standardised assessment of the quality of student work.

1850-1862 Fisher Scale Book and numerical approach

The next regularly referenced innovation in quantified assessment is Reverend Fisher's Scale Book. Procedures to manage the quality of student work within individual schools, in the middle 19th century, are not well described in the literature of the time. One exception is a brief paper written by Reverend Fisher, Principal of the Greenwich Hospital School in the U.K. *c.*1862. Fisher was encouraged by Chadwick, President of the statistical section of the British Association for the Advancement of Science to document his assessment processes. Fisher's paper subsequently became a widely referenced 19th century example of an approach to teaching quality, student assessment and scale development (Ayres, 1918, cited in Cadenhead & Robinson, 1987; Cadenhead & Robinson, 1987; Haertel & Herman, 2005).

Fisher had been a naval officer, chaplain and astronomer in the 1820s and 30s, accompanying the Buchan and Parry expeditions to the Arctic in a role similar to that of Darwin and Huxley in their own scientifically formative voyages (Darwin, 1860; Huxley, 1936). Fisher's systematic scientific observational background might help to account for his quantitative approach to classifying the quality of student work.

Fisher presented his process in a paper to the 32nd Meeting of the British Association for the Advancement of Science in October 1862 (Fisher, 1862) with the title *On the Numerical Mode of Estimating Educational Qualifications, as pursued at the Greenwich Hospital School*. Fisher recognized the utility of a numerical representation. It was not only an efficient and information-rich mode of recording "easily referred to at a future time" but it afforded "also the means of determining the average condition of a class or school, as regards each subject of instruction and also the whole amount of educational work done." (Fisher, 1862, reprinted in full in Cadenhead & Robinson, 1987, p. 17). He was able to adapt the data for graphical representation, plotting the "mean values of the various educational qualifications of the boys at the completion of their education from 1850 to 1862 at each quarterly examination" (p. 17).

Chadwick (1864) reported an interview with Fisher who explains the rationale for the development of the arrangements put in place at Greenwich. "We had no records of results and it was to supply the deficiency that the numerical method was devised by me. The teaching was of a very inferior character." (Fisher in Chadwick, 1864 p. 263) Fisher developed his Standards Scale book in response to the perceived inadequacy of descriptive terms good, bad, indifferent and so forth, which he saw as subject to "various and somewhat uncertain interpretations, … arising from the fact that no recognized standard or fixed scale has hitherto been employed in assigning the absolute and comparative values of such expressions" (Fisher, 1862, reprinted in Cadenhead & Robinson, 1987. p. 16). In his own words his intention was to "refer such elementary attainments to standards which approximate to a permanent character to numerical equivalents for such terms, to afford more accurate and precise meanings than the words allude to, and at the same time [provide] a more concise mode of registration, combined with the means of integrating or expressing the sum-total of any number of results." (p. 16). The method used numerals one to five to denote the standard of work.

The scale-book contained examples of varying degrees of proficiency with a numerical value for each. For example, to "determine the numerical equivalent to any specimen of writing, a comparison is made with various standard specimens of writing contained in this book, which are arrayed and numerically valued according to the degree of merit" (Fisher, 1862 in Cadenhead & Robinson, 1987. p. 16). This anticipated by 50 years a similar process developed by Thorndike in 1910. The scale's highest value was 1, lowest 5. Scale points at a

quarter of a division were used to denote intermediate values, allowing in all 17 scale positions from 5 to 1[5].

The scale-book also included spelling, mathematics, navigation, scripture knowledge, French, general history, chart drawing and practical science; in these cases providing questions in each subject, to serve as types of the difficulty and also the nature of examinations. The scale-book did not include reading, characters and natural talents "where the usually received interpretations of the words 'good', 'bad' etc." (Fisher, 1862 in Cadenhead & Robinson, 1987. p. 17) were deemed adequate. Rev. Fisher did not regard reading performance as needing to be scaled.

The system had utility in the Greenwich Hospital School but seems not to have spread too far into the English school system. How the teachers in the school might have felt about the process is not reported. The scale book was systematic and numerical and is preserved for posterity through Chadwick's interest in a quantitative approach to documenting education.

1892-1908 Rice's educational surveys and Stone's enhancements

In the U.S. the next widely acknowledged developments in the evolution of educational assessment are the surveys of Rice. His initial survey of 1892-93 was observational, richly described and included student work collected from his visits (Rice, 1893). He observed the general instruction independent of topic, looking in particular for 'scientific teaching'. His methodology changed for subsequent surveys in spelling in 1897 and arithmetic in 1902 (Rice, 1913). In these latter surveys he collected written responses and provided statistical analyses. His findings confounded the administrators and teachers of the day. Schools providing 15 or 20 minutes of spelling daily did as well as those providing 40 or 50 (Rice cited by Thorndike, 1914a, p. 293). While Thorndike criticized the methodology (lack of recognition of the importance of the difficulty of words chosen to be tested) by implication he accepted the general drift of Rice's findings and acknowledged the importance of Rice's data collection approach (Thorndike, 1916a, p.5 and p. 9).

---

[5] Scale note: It is likely that Fisher assumed an equal interval scale. This thesis considers the utility of the logit (log odds unit) as a unit of measurement for learning development. Using assumptions about skill development over an extended period of time (5-7 years) and based on parameters from current test measures (Hungi, 2003) it is possible to estimate that a scale increment of a quarter of a scale unit ($1/17^{th}$ of the full scale) was likely to have been of the order of 0.2 logits in current terms (Estimated 3.5 logits from value 5 to value 1, divide by 17 units). Whether this level of precision is practical is an issue addressed later.

Stone, following a survey approach similar to that of Rice, collected data about arithmetic in 1908. He developed two arithmetic tests, one addressing 'fundamentals', that is the four operations of addition, subtraction, multiplication and division; the other addressing 'reasoning', the solving of word problems with relatively complex arithmetic and logic.

He acknowledged his debt to Rice (Stone, 1908, p. 96) but believed he had made improvements on Rice's approach. These were improved instrument and research design and improved methods of securing and handling data. The improvements in the gathering and handling of data were "chiefly those of refinement, and they could hardly have been planned for without the benefit of Dr. Rice's and other pioneer studies.' (Stone, 1908, p. 96) Stone was of the view that reasoning and fundamentals were different abilities "and should be so measured" (1908, p. 96). A number of other refinements were made to address, amongst other matters: time allowances for the test, test-room procedures, scoring, access to the data and computations (i.e. data analysis) by other researchers, disassembling the fundamentals into the four basic arithmetic operations (addition, subtraction, multiplication and division) and the use of correlation coefficients (Stone, 1908).

Stone's analysis was comprehensive, considering that all the tabulations and computations were completed by hand. He presented his data as aggregations of scores or aggregations of errors made. His major conclusion was: "Probably the truest single expression of the findings of this study is summed up in the one word, diversity." (Stone, 1908, p. 90) He noted that the within-system variability was greater than the between-system variability. In his view the greatest need identified by the research was the promulgation of standards of achievement. This need is understood today as a need for benchmarks. "That the great variability herein shown would exist if school authorities possessed adequate means of measuring products is inconceivable." (Stone, 1908, p. 90)

Of the surveys published to this time, Stone's was the first to show an appreciation of the issue of item difficulty. Stone's concern with item difficulty was twofold; he wanted to present items in tests in order of increasing difficulty and also to assign a weighting for more difficult items in the data analysis. He recognised the importance to his analysis of weighting the more difficult items. In this he was possibly influenced by Thorndike who gave "guidance in executing the statistical phases" (Stone, 1908, p. 5). He considered two options for weights. One related the proportion successful on the most difficult item (12% correct) relative to other items including the easiest item (94% correct), generating a wide range of weights with a direct relationship to the log of odds ratios (author analysis). His second option restricted the range to a maximum of 2 for the hardest items. As a result the weights for a selected a set of items were unity (i.e. weighting=1 for the easiest set), a hard set with a

maximum weight (2) and an intermediate set of 3 items with weights scaled at independent values between 1 and 2, on what can be established as a log odds ratio basis (author analysis).

Stone applied the weighted transformation to the total score for a system rather than to each student. As a consequence of the weighting, school systems had their aggregate score (items correct per 100 students) weighted in very approximate proportion to the log-odds transformation of the more difficult items (author established not detailed here). The resulting score per system was stretched for higher performing systems relative to lower performing systems[6], very crudely applying one principle of the logistic transformation used in the Rasch model. For its time this was an important and prescient insight by Stone but he had no reason to see the scale of his reasoning items as having a value beyond its contribution to stretching the score of high performing systems relative to lesser performing systems. His work, however, encouraged others to attend to what learning was actually occurring in schools.

1909-1911 Courtis and the influence of Stone

In 1908 Courtis was head of Science and Mathematic Department of the Liggett School, Detroit. Immediately after the publication of Stone's analysis he applied the Arithmetic tests to all students in the school (Courtis, 1909a). He published the results of a series of tests and the subsequent refinements he made to the testing process, in instalments, in *The Elementary School Teacher* (Courtis, 1909a, 1909b, 1910, 1911a, 1911b, 1911c). His first instalment applied the test unchanged but varied the scoring process. He did not adopt the weighting system applied in the reasoning test, thus moving away from the Stone insight of increasing the distance of higher scores from lower scores.

Initially Courtis's interest was in seeing how his students compared with Stone's data but, unlike Stone who tested only grade 6A, he was also interested to see the effect across the whole school. Thus he applied the test, unchanged, to Grades 3 through 13, testing 218 students in all. He described his reason as establishing standards for judging the success of a reorganisation of the mathematics course in the school. He also declared an interest in tracking

> … the development of ability in arithmetic from the primary grades through the high school. Such tests, repeated at frequent intervals, would … make standardization of

[6] The highest scoring system had an unweighted score of 748, a score of 914 with the preferred weighting, and 1266 with the original almost odds ratio scale. The preferred weights increased the base score by 1.21, the original weights by 1.69. The relative increases in scores for the lowest scoring system (341) were factors of 1.01 and 1.11 respectively. (Stone, 1908, p. 98, Table XXXVIII)

> yearly work possible, would show exactly the place, manner, and amount of development of any particular ability, and would give a rational basis for the estimation of the influence exerted by any method, material, or teacher. (Courtis, 1909a, p. 58)

While it is possible that other teachers of the time were interested in documenting and understanding the development of mathematics learning, Courtis appears to be the first to publish the trend across a school, triggered by the publication of the Stone report. As unfolds below, he had a significant impact on the teaching and assessment of arithmetic in a large number of schools as a result.

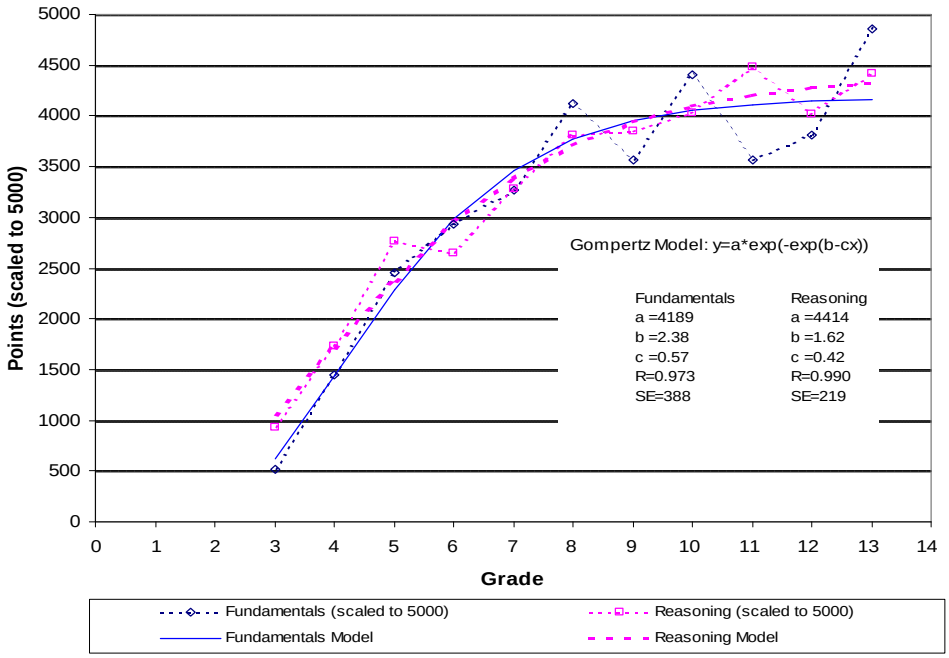
Courtis initially took a teacher and school perspective on the use of tests, and applied a new technology to understand learning at his school. Ultimately it led to a series of insights about the range of performance of students within classes and across classes, in what would appear to be the first published cross-sectional analysis of a school using a common quantitative assessment process. The test was clearly inappropriate for many of the younger students but some were able to complete the additions and solve some of the reasoning problems. Instead of using average scores (or median scores) for each grade, he presented the total aggregate score for 25 students per grade, requiring an adjustment to most cohorts to convert them to the 25-student standard.

He recorded the number of examples attempted, the number correct, the accumulated point value for the items attempted and the accumulated point value for correct steps taken, even if the final result was incorrect; a part credit scoring process. As an example, a question such as "divide 278542 by 679 is made up of 6 additions, 7 subtractions, 9 multiplications and 4 divisions" (Courtis, 1909a, p. 62) leading to a maximum possible score of 26 points. Courtis graphed the data to help the reader appreciate the apparent pattern of change with grade, showing an early growth of skill by grade view of arithmetic learning, along with variability across grades.

Courtis observed a pattern of alternating strengths in either reasoning or fundamentals by grade, one appearing to be stronger than the other at particular grades. He formed a view that arithmetic skills might be more fairly represented by a composite score, where the two aspects might balance each other out (in contrast to Stone who saw merit in keeping the aspects separate). Recognising that an unweighted comparison of fundamentals and reasoning grade point scores would be biased in favour of the more highly scored fundamentals (just under 5000 points per class) versus reasoning (just under 800 points per class) in combined raw scores, he rescaled both to a common scale of 5000 points (based on an assumption of a possible maximum fundamental score of 5000 for Grade 13); in what he called equating.

Figure 2.1 illustrates the result, redrawn from Courtis's (1909a) original data. He added hand drawn curves of visual best fit. Courtis showed (Figure 2.1), that grade variations in the relative skills of fundamental operation accuracy and success in reasoning problems, generally develop together. Although at any particular point one skill might appear stronger than the other, it may be an artefact of measurement error.

Figure 2.1 Points Scores for Correct Steps-Fundamentals v Reasoning



Note: redrawn from Table 11 of Courtis, 1909a, p. 71.

Smoothed trajectory curves hand drawn in original. Gompertz expression applied in example.

In lieu of Courtis's hand drawn curve, a computer-assisted curve-fitting process is applied to the data series for fundamentals and reasoning. The curve describes visually (and mathematically) the average trajectory of the development of these two skill sets, as reflected in Courtis's data, by grade. Using *CurveExpert* (Hyams, 2001) a range of curves can be tested for fit. The best fit is obtained with a Gompertz[7] model (Gompertz, 1825), satisfyingly appropriate as this model was ultimately chosen 17 years later, in 1926, by Courtis as the most likely model to describe controlled growth (Johanningmeier, 2004, p. 205). The curves establish a diminishing rate of growth in score points as Grade increases.

[7] The sigmoid model that best fits the Fundamentals data is a Gompertz model (SE= 388.3, R= 0.97). This fits slightly better than a logistical or MMF (Morgan-Mercer-Flodin) model. The sigmoid model that best fits the Reasoning data is also a Gompertz model (SE= 219.2, R=: 0. 99). This fits slightly better than a logistical or MMF model. See Appendix 5 for further information relating to the use of *CurveExpert* and the Gompertz model generally.

The idealised (i.e., fitted, smoothed) paths of development for both skills graphed in Figure 2.1 appear to follow, superficially at least, similar and almost coalesced trajectories for large segments. Whether this relationship would be sustained if the two aspects were equated using, say, procedures based on the Rasch model and a logit scale applied in lieu of rescaled points, cannot be determined. The means of raw scores for grades are assumed to correlate very highly with the mean of the Rasch model transformation of the individual scores, given the dependence on total scores in the Rasch model. The resultant increase in the spread of transformed scores is not likely to change the relationship of the two trajectories markedly.

Courtis's work was important in the development of quantitative educational assessment from the school perspective, as distinct from the system approaches of Rice and Stone. He appears to be the first to consider the time dimension of development using school grade. The publication of his analyses of data from the Liggett School for Girls, Detroit continued through 1910 and 1911 (Courtis, 1909a, 1909b, 1910, 1911a, 1911b, 1911c). Having used Stone's approach initially, Courtis moved to design his own instruments, which became popular very quickly.

By 1911 he had developed his own series of eight tests, the *Courtis Arithmetic Tests Series A,* and provided 30,000 sets throughout the US, England and Germany (Courtis, 1911c). He recognised the value of reference data being provided back to schools and aggregated 9000 individual scores across 14 grades from his tests to show patterns of typical development by grade (Courtis, 1911c). By 1913 with over 55,000 cases analysed for grade averages, he described the purpose of his tests as enabling the study of arithmetic abilities. He ultimately designated his initial test as Test 7 and designed simpler tests to lead up to this level of difficulty. From his analysis of the mistakes made by students he identified the "necessity for diagnostic tests of the simpler component abilities … and tests Nos. 1 to 5 were constructed" (Courtis, 1913, p.329).

By 1914 Courtis had broadened his view of useful data for observing the student and classroom. Under the slogan of "Measure the efficiency of the entire school, not the individual ability of the few" (Courtis, 1914, p. 380), he packaged a range of tests for English language development, covering handwriting quality, legibility and rate, composition generally, punctuation, spelling and syntax along with tests of memory. The handwriting tools were adapted from other authors (Thorndike, 1910 and Ayres, 1912) but other elements were of his own design.

In discussion of his approach to standard tests in English, Courtis anticipated a version of Vygotsky's Zone of Proximal Development (Vygotsky, 1978, p. 86). Based on his observations in English and Arithmetic, Courtis argued "it is not possible radically to change

the efficiency of present methods *until the actual work assigned to each pupil is based on his measured needs*" (1914, p. 392, italics in original). Courtis further explained that his tests "will furnish objective standards that will serve as goals for the guidance of teachers and pupils, and as a means of detecting the peculiar weaknesses of individuals." (1914, p. 392)

He concluded that the spread of achievement of reading was very similar to that of arithmetic and that he "expects to find that the same general causes operate to prevent success, that the same factors determine efficiency, and that the same changes in methods of teaching will prove effective." (Courtis, 1914, p 390)

He also anticipated some consequences of the Piagetian insights on stages.

> It has been a puzzling fact of teaching experience that ability to reason and ability to be exact in abstract work seldom go together. He is inclined to believe that there is a psychological principle at work, which, if known, would solve more riddles than one in educational procedure. Whatever the explanation, statistical proof of the fact is given here … Accuracy gradually decreases through the grammar grades [i.e. elementary/primary] and increases through the high-school grades at about the same rate. If this result is confirmed by future tests, there is an important lesson here. If inaccuracy in grades 7, 8, and 9, is due to some natural cause outside of arithmetic proper, to insist on accuracy or to spend much time in working for it may be not only wasteful, but harmful. (Courtis, 1909a, p. 73)

Having explored arithmetic, and anticipating his next phase into English language, Courtis speculated in 1909 that subjects taught over successive grades could be understood developmentally or longitudinally. He saw his tests as providing "a connective thread of growth in the fundamentals of the subject that will produce a unity that is sadly lacking in all present pedagogical effort." (Courtis, 1909b, p. 199) To understand student development over a broader time spectrum, Courtis recognised the need for observation and analysis that could connect across repeated tests and across the grades of the school. Single tests within grades, offered little if any insight to student learning development over time unless the tests could be connected to each other in some way.

In 1916 he reported 455,000 tests were sent out in one 12 month period (Courtis, 1916). His Teacher's Manual (Courtis, 1917) for the use of the practice arithmetic tests made clear the link he saw between testing and classroom practice. He offered advice on the efficient use of his tests to select who in each class should be involved. He targeted his support to the level of each student and encouraged teachers to adjust "the general method to … local conditions" (Courtis, 1917, p. 2). He summarised the steps in the use of his approach as

> a. Measure your class to determine the initial ability of its members.
> b. Eliminate from the drill class those who have (or reach) standard ability.
> c. Give to each of the other members drill upon those lessons where drill is needed.
> d. Permit each individual to practice in his own way and to grow at his own rate.

e. Give exactly the assistance needed to each child that fails.
f. Measure the efficiency of your teaching. (Courtis, 1917, p. 2)

His strategy for the use of measurement fits with the spirit of the arguments herein about how a teacher, having the best estimate of a student's learning status, might be expected to respond today, all-be-it that the range of supports should now be wider.

Courtis will be revisited briefly in Chapter 5 where further development of his concern about the relationship of learning growth with time is considered. At this stage the key insights from the early Courtis work are:

> teachers can be supported to use standard scientific processes to understand what is happening in individual student learning and within cohort learning;
>
> observing a range of aspects of learning developing together over grades enables their development with time to be understood; and
>
> many teaching interventions seem not to influence the rate of development yet students eventually improve.

Courtis's work could be labelled evidence based in today's terminology but he later became sceptical about tests. He withdrew his tests from the market in 1938 after twenty million copies had been sold because he "discovered they did not measure what they were supposed to measure" (Johanningmeier, 2004, p. 205). Courtis argued that repeated measures of the individual's "progress in terms of his own growth curve" were more useful than external norms and that growth was "cyclic in nature" (Johanningmeier, 2004, p. 206). His work provided an important impetus in encouraging teachers to observe the development of their students from a scientific perspective. (In a Frankensteinian escape, the tests took on a life of their own – independent of the intention of their creator.)

1910 Thorndike and the handwriting scale

Courtis's exploration of arithmetic coincided with an explosion in the range of assessment tools. One example was Thorndike's handwriting scale (Thorndike, 1910), which Courtis adopted into his English language assessment suite. A set of examples of handwriting was provided, each with a scale value that placed it on the scale at equally spaced intervals of quality (in Thorndike's view).

Thorndike (1910) published the scale in the *Teachers College Record* but was a little vague about the exact process of derivation. From pages 4 to 7 of the article it can be inferred that he followed the following steps: He selected a thousand examples of handwriting, many supplied by Rice, that were then rated into about 11 groups by 40 judges. As a result, each

example achieved an average score over all the judges, in the range 1 to 11. Thorndike selected examples close to the averages of 1, 2, 3, 4, etc. as his final examples and argued that they were about 1 unit apart in improving handwriting merit. To check the selected examples, he followed a second process of equally often noticed differences, where the judges compared the selected examples in paired comparisons. He explained "only if differences are not always noticed can we say that differences equally often noticed are equal." (Thorndike, 1910, p. 6) On this basis an example can be described by the percentage of judges who found a<b, a=b or a>b. Thorndike required his examples to show a separation of about 75:25[8] to justify a one-unit difference in the scale position. He also argued that 10 to 15 examples adequately spaced would be sufficient for a teacher to place an example of a student's handwriting at either of (or between) two scalar examples.

Thorndike considered the issue of where to place the lower and upper reference examples and argued that weaker and better examples than those on the scale of interest should be provided.

> The scale extends in actual samples by children from nearly the worst writing of fourth-grade children (quality 5) to nearly the best writing of eighth-grade children (quality 17). Quality 7 is nearly the worst writing of fifth-grade children.
>
> The scale includes a sample of a copy-book model which is rated by competent judges as of approximately quality 18, two samples of fourth-grade writing which are judged to be approximately of qualities 6 and 5, and a very bad writing, artificially produced, which is rated by competent judges as of approximately quality 4. The scale thus extends from a quality, better than which no pupil is expected to produce, down to a quality so bad as to be intolerable, and probably almost never found, in school practice in the grammar grades.
>
> If one had a finer scale, its use would give but slightly more accurate results, and would require more practice and more time. (Thorndike, 1910 p. 8)

The degree of fineness of the scale is considered again, later in the chapter, in the work of Hillegas. Thorndike argued the scale was necessary to be able to measure differences in the quality of handwriting. In a variation of the oft reported aphorism he claimed the

> …history of the judgments of the merit of handwritings supports the claim that if a number of facts are known to vary in the amount of any thing which can be thought of, they can be measured in respect to it. Otherwise, I may add, we would not know that they varied in it. Wherever we now properly use any comparative, we can by ingenuity learn to use defined points on a scale. (Thorndike, 1910, p. 69)

Thorndike saw the handwriting scale as a reference, something to which a teacher might need to refer in the assessing of the quality of any student work, and that through use, the scale

---

[8] The natural logarithm of 3 (75/25) is 1.09, that is approximating one logit, indicating that his scale (based on judges) has an approximate relationship of one unit to one logit.

would become internalised (that is referenced often enough to confirm an ongoing personal judgement calibration). Thorndike argued that the scale could be used as a mental standard, in the same way as an estimate of length might be made without using an actual ruler but by a tacit understanding of length units. He envisaged users of the scale having a stored impression of the quality examples.

This is the essence of the concept of teacher judgment of educational development considered in this thesis: applying a method that gives numerical substance to qualitative descriptions or categories. A framework is developed, used, internalised and the teacher's judgements becomes calibrated to the scale, with infrequent checking back to the original calibration examples.

1912 Thorndike's concept of scaling

Thorndike (1912), addressing the Harvard Teachers' Association, argued that educational science was able to apply the same process as led to physical scales to a wide range of educational developments. "Scales, graded standards, by which to report knowledge of German, ability to spell, skill in cooking, original power in mathematics, appreciation of music, or any educational fact you may think of, are now where the thermometer, spectroscope, and galvanometer were three hundred years ago - they do not exist." (Thorndike, 1912, p. 291) Using the scale for weight (in his example, in grams) he argued for four elements of an ideal scale:

> A series of perfectly definable facts; …
> Each amount is a different amount of the same kind of thing; …
> Differences between any two amounts are perfectly defined in terms of some unit of difference; …
> The zero point is absolute, it means 'just barely not any' of the thing in question. (Thorndike, 1912, p. 291)

He then described a range of educational development areas where "it is an easy task, theoretically, for educational science to take … vague, ambiguous statements of common-sense and refine them as physical science has in the past refined similar measures in the case of physical facts." (Thorndike, 1912, p. 292) Drawing on the concept of difficulty he explained that "in the case of spelling, we can define a point on the scale as the ability to spell words as hard as, but no harder than, 'a' and 'go', or 'wish' and 'touch,' and so on to 'millinery,' 'development,' or words of any difficulty we choose." (Thorndike, 1912, p. 291, commas as in original.)

Thorndike went on to explain the method of equally noticed differences, derived from Galton and Cattell and as used above in his handwriting scale. He used as his example the composition scale under development by Hillegas (of more, later) as an example of how

passages of writing of varying qualities could be rated by judges to create a scale, similar but more complex in its concept, than his handwriting scale.

Thorndike anticipated at least "two or three objections" (p.299) that teachers and administrators might have with a scaled approach:

> …the good old adjectives are enough for educational work
> …the common-sense judgment of a first-rate man without these units and scales is better than the action of the stupid man or incompetent man, with them.
> …the personal, spiritual work of education - the direct human influence that the pupil may get - is not in the domain of exact science. (Thorndike, 1912, p. 299)

He countered the first objection (existing adjectives good enough) with the assertion that for the kind of person making that objection, the use of the vague descriptors approach would suffice. As to the second objection, he acknowledged that a knowledgeable person without the scale might make better judgements than a "stupid man or incompetent man" with it but that it was the work of science "to get good work done by those of us who are rather mediocre". (Thorndike, 1912, p. 299) To the third objection he argued that the benefits of measurement and precision were not in conflict with more ethereal matters. "Mothers do not love their babies less who weigh them. We do not serve our country less faithfully because we take its census" (p. 299). While not exhaustive of the arguments of the 21st Century, his broad sweep covers some of the current concerns that measurement and judgment evoke.

1912 Hillegas: judging the quality of prose

The next application of Thorndike's scaled (holistic) judgment approach as applied in handwriting was that by Hillegas (1912) who, as a student of Thorndike, created a scale for the judgment of the quality of prose composition. The process is instructive in the labours taken to achieve this scale.

To start he acquired 7000 composition examples from "various sources and represent a definite attempt to obtain particularly the very poorest and the best work that is done in the schools" (Hillegas, 1912, p. 22). From these he selected 75 examples, supplementing the upper and lower ends of the scale with manufactured examples. The lower end samples were created by adults consciously trying to write poorly and the upper end from the youthful writing of Austen and the Brontes. Thus he started his calibration with 83 examples. The examples were typed with all characteristics retained (misspellings, punctuation etc.) to avoid the quality of the handwriting complicating the assessment, and then duplicated. In the first phase 100 judges were requested to rank the 83 compositions from worst to best and signify the order by numbers 1 to 83, or fewer if ties were required. Only 73 judges were able to follow the instructions correctly.

From the first responses Hillegas selected 23 compositions based on a process that selected the examples on the basis of steps in merit. He achieved this by selecting the poorest (about 95% of the judges agreed on this case), and the next two weakest examples. The balance of the examples were selected by finding the first case relative to the previously selected case that 75% of the judges had selected as better. Two gaps between the three weakest examples in the range were filled again with artificial samples. His aim was to create a scale that met Thorndike's ideal with a well-ordered set of examples spaced at exactly one Thorndike unit.

A revised set of 27 scripts was then sent to over 100 additional judges (teachers, authors, literary workers). In the second iteration the task was to create an ordered pile with the best at the bottom, the worst at the top, to be returned securely fastened. The first 75 to return were tabulated. Thorndike meanwhile had obtained 41 additional judgments from "individuals who were especially competent to judge merit in English writing", which were tabulated separately so their rankings could be used a "check on the others" (Hillegas, 1912, p. 40).

Additional judgments were also solicited from the general science community through an article in *Science* of June 1911 by Thorndike (Thorndike, 1911), which included only 8 of the original 27 cases for ranking, plus the latest two additions. The fate of these responses is unclear. Hillegas acknowledges 515 sets of responses overall, 202 of which were used in his analyses.

The result of all this labour was a set of 10 examples for use in the judgment of English compositions, the Hillegas Composition scale. It appears to be the first scale after the Thorndike Handwriting scale, to be offered as a tool to provide teachers a method to calibrate their judgment of composition merit. The items had values of 0, 183, 260, 369, 474, 585, 675, 772, 838, and 937. The scale units were 100 times the 'raw' unit that came out of applying the Thorndike scaling process. As described earlier this process assumed a unit of 1 (or 100 on the scale above) for a case where exactly 75% of judges rated a case as superior to a lesser case, based on Thorndike's use of the Median Deviation[9]. Thorndike (1916a, p. 228, Table 59) created tables to convert the difference in percentage of judges, working to two decimal points of precision, which one assumes Hillegas had used to look up the values. The implied precision alone may have been sufficient for many teachers to be sceptical about its utility.

---

[9] The Median Deviation is the median of the set of absolute deviations from the Median. It has a regular relationship to the Standard Deviation, with the constant dependent upon the type of distribution. For a normal distribution the SD is approximately 1.486*MD (now usually referred to as the MAD -Median Absolute Deviation). (http://en.wikipedia.org/wiki/Median_absolute_deviation)

The scale's purpose was to provide a ruler for a 'holistic' quality judgement. The actual characteristics of composition merit were not teased out; that is the composite elements of quality were not identified or made explicit. This led to some criticisms of the scale. Hillegas devoted his closing paragraphs of his 1912 article to defending his choice of multiple types of writing in one scale, requiring a holistic judgement. His defence of his approach was that

> people actually did do it. Of the four hundred and fifty people who have judged these samples not more than three have offered any objection on the score that they could not compare the samples. (Hillegas, 1912, p. 55)

Others were enthusiastic in their promotion of the Hillegas scale for system and classroom use. Abbott, of Teachers College, Columbia University, distributed copies of the Hillegas-Thorndike Scale at the 1916 conference of the National Council of Teachers of English (National Council of Teachers of English, 1917). He reported that through the use of the Hillegas scale the variations in the markings of freshman essays were greatly reduced, but that the markers felt the need of a scale to distinguish between form and content. It was reported that the scale had been used in Salt Lake City where fourth-grade pupils attained an average of 29 (the last digit had by now been deleted); fifth grade, 31; sixth grade, 38; seventh grade, 44; and eighth grade, 54. This was interpreted to show that steady progress in composition merit was being made. Cross (1917) advocated scales generally for the classroom teacher and believed the Hillegas scale was effective. "It would seem that there is too much room for individual opinion in judging here; but in the experiments I have made with the scale, having a number of persons read the same composition and then grade it by the Hillegas scale, the results were much more nearly uniform than I expected." (Cross, 1917, p. 188)

Thorndike (1913) considered the issue of errors of judgment using the scale, and speculated on the likely behaviour of teachers using the scale. He considered that initially errors would be large but that they would "diminish with practice in using such a scale and with improvements in the scale itself". With practice, he believed, errors would be "smaller than the errors now made by teachers in grading paragraph-writing for general merit" (Thorndike, 1913, p. 556). He argued that the reason for the errors being smaller was that a teacher, in grading a composition for general merit, used a subjective, personal scale of values that "cannot, on the average, be as correct as one due to the combined opinions of a hundred or more judges who are on the average as competent as he is." (Thorndike, 1913, p. 556) Hillegas's scale, he claimed, eliminated the errors due to the personal scale altogether and with enough practice with it would probably decrease the errors of comparison.

Thorndike (1913) and Trabue (1917) provided additional or alternative items for the Hillegas scale (creating the Thorndike Supplement and the Nassau County Supplement). In principle these samples were use in an attempt to maintain equivalent difficulties to the original scale

so that the scale was preserved. Trabue (1917) argued that the original scale, while helpful, had some deficiencies: artificiality of the lower examples, the brevity of the examples, a need for examples to be more similar to the type written in Nassau County, and, finally, some indicative data of standards by grade were desirable.

These early prose judgment scales, while controversial and imperfect, confirmed that it was possible to quantify classroom phenomena on the basis of teacher judgment applied to authentic student work, as long as teachers were provided with a reference frame or scale. Subsequent scales developed in the same period were more specific, that is targeted to very particular skills, for example, punctuation. There was also a trend towards conventional test schemes, those concerned with performance at the point of testing for selection or grouping purposes. This test score tradition is described by Engelhard (1991b, 1992b) as developing strength from the work of Wood, also a student of Thorndike. Wood was "a driving force behind the measurement movement of the 1920's that replaced essay examinations with multiple-choice items" (Engelhard 1991b, p. 146).

1913 Criticisms of scaled approaches from researchers of the period

While the scales had their advocates, they also had their critics (Johnson, 1913; Thomas, 1913; Learned, 1913; Neilson, 1913; Thurber, 1913; Holmes, 1913).

Johnson (1913) set three different groups (N=42; 16; and 5) the task of applying the Hillegas scale to eight composition examples. The range of variation from the lowest to the highest allocated values for any one item, averaged over all examples, was 3.7 Thorndike Units. The mean scale values for each item from each of the three groups, however, were close. The mean ratings for each example for each group correlated with the other two groups at between 0.98 and 0.99 (author calculation) indicating that while the within group variation was large, the orders of the average scores in each group were very consistent. Based on the upper and lower range values (the only exact data points reported) it is possible to observe that judges assigned values outside those of the examples in 65% of these cases. Judges used the scale as Hillegas and Thorndike expected and were not limited to the specific example values.

Learned (1913) reported an investigation where 50 papers were graded by 15 teachers. Initially the papers were graded using a percentage scale, then a month later with the Hillegas scale. The spread of values was reduced to 75% of the original range for all judges when the Hillegas scale was used, and to 56% of the original range for the 9 judges closest to the median. Thurber, in the same leaflet, (Thurber, 1913, p. 7) took a strong negative position, which on the face of it could have been a tongue in cheek argument for scales, though from the context this was unlikely to be so.

> The most baneful effect of the use of scales is that they inevitably make theme correcting more objective, and less subjective; the teacher's attention is at once focused upon the paper and not upon the boy who wrote it,—upon abstract qualities of writing, not upon personal qualities of the writer. The Hillegas Scale, as any number of better scales, used ideally, would make it possible for any English teacher in the country to correct and mark papers exactly as well as the teacher for whom those papers were written. Such a thing, on the face of it, is absurd. (Thurber, 1913, p. 7)

The Thurber comment highlights one of the tensions in the use of scales and judgement. The assessor is focussed on the merit of the student product. In principle, the same product should be judged by all teachers to be positioned at approximately the same place on the scale. Thurber's reason for the assessor to take into account "the personal qualities of the writer" in the judgement is not clear.

The Hillegas scale was still being used in educational research as late as 1940 but not as a classroom scoring tool or scale but as a method to identify text pieces of differing quality (Hinton, 1940). The process to develop the original scale was comprehensive and perhaps more complex than it needed to be. The Hillegas scale had some utility but it did not survive. It illustrates the principle that reference examples can be used to provide a score to a piece of writing and that values selected by assessors are not limited to those of the examples. Assessors used the examples as a scale and estimated values in between scaled examples. The principle of holistic marking of essays has been retained in modern US testing processing, although with less refined scales and, on occasions, less sophisticated markers than experienced classroom teachers (Farley, 2009).

1914-1916 Thorndike's scaling for reading

Thorndike (1914b, 1915, 1916b) applied his scaling approach to two elements of reading: reading words adequately to categorise them, and silent reading of passages of increasing complexity and then answering comprehension questions of varying complexity. Gray, another Thorndike student, developed at the same time a set of reading passages for reading aloud covered in the same article (Thorndike, 1914b). Scales were developed for these three aspects of reading development.

Thorndike assumed invariance of item difficulty over time and across locations for all items. He acknowledged however that there were local variations in difficulty. Words unfamiliar in one region may be commonplace in another, though he argued that these variations were usually exceptions and had only a small impact on his general approach. This general invariance of the difficulty structure of sets of items is a fundamental requirement if any scale of learning is to be feasible.

Thorndike's empirical approach was time intensive to develop and thus beyond the resources of classroom teachers. The product, however, was readily applicable in the classroom and by the 1915 version able to help assess the skill development of individual students. A benefit was that "the results will be readily comparable with thousands of others obtained with other classes by other supervisors, and will be at once understood by anybody who knows the scale—a most desirable feature" (Thorndike, 1914b, p. 14). His assessment approaches were eclectic, encompassing test items for students (words of established difficulty) but also holistic judgements (the Thorndike writing scale or the Hillegas scale). He was not averse to this use of observer judgement, believing he had processes to scale judgements as well as word difficulty.

1916 Trabue's completion test

Another of the scaled tests, the Trabue Completion Test was the subject of further investigation by Engelhard (1984) and is described briefly as it indicates how well scales of difficulty were being applied. The test itself, based on the Ebbinghaus-developed idea of filling missing gaps in text, was used as an indicator of the ability to think about words and language. Engelhard (1984) considered the data reported by Trabue informative about the scale structures developed by Thorndike, and an alternative approach to scaling developed by Thurstone.

Trabue's test began with sentences so simple that a large majority of the second-grade pupils were able to complete them correctly, and finished with sentences so difficult that only a small percentage of freshmen in college could complete them; that is, he had a concept of ordering the test by the difficulty of items. With an interest in measuring progress from year to year or from grade to grade, Trabue established empirically the difficulty of each sentence by trialling the incomplete sentences with thousands of public school children. From the results he developed four approximately equal scales, each scale consisting of ten sentences. He explained that by measuring ability at the beginning of a year with one scale and then at the beginning of the next year with an equivalent scale, it would be possible to determine the amount of progress made by a class or by a child during a year (Trabue, 1916, p. 88).

Trabue explored partial credit scoring as part of his analysis of the performance of the test. He tried initially six grades of quality (5-4-3-2-1-0) in the completion of a sentence. Trabue found that nothing was lost by "simplifying the scoring still further, giving two points credit for each perfectly completed sentence, one point for each sentence completed with only a slight imperfection, and zero for any sentence omitted or imperfectly completed". This "had benefits in efficiency of marking with no loss of information for scoring students" (Trabue, 1916, p. 87). He viewed items as being linearly scaled in difficulty, with the point where

students were unable to add an appropriate word being the measure of current student ability, implicitly seeing students and items on the same scale and also implying item invariance (Trabue, 1916).

Item Difficulty and the key link to educational measurement

Were the items spaced along a continuum of difficulty in a fashion comparable to that might be established today?

Engelhard (1984) acknowledges that the problems and issues in psychometrics have not changed much since early 1900 and were well considered by Thorndike, and later Thurstone. The conditions necessary for objective measurement as described by Rasch model advocate, Wright (1968, cited in Wright, 1977) include: the calibration of measuring instruments must be independent of those objects used for calibration, and the measurement of objects must be independent of the instruments used. These conditions require the objects/items used in calibration of scales to be invariant in relative difficulty across samples or occasions.

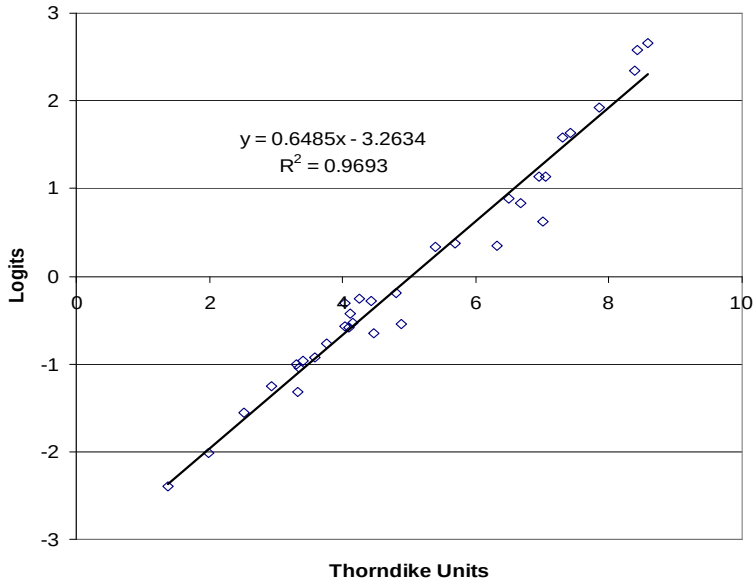
Thurstone (1925, p.433) commented on the inadequate discussion of the "assumptions and the logic of … scale constructions" and proposed a new approach to scale construction. He illustrated his process with a scale he developed for the Binet test questions, re-analysing Burt's data from 3000 London school children. His analysis, also based on standard deviations, established a scale of order of items for the Binet test questions, with an origin at the mean of Binet Test intelligence for $3^1/_2$ year olds. The result was the first published item map. It illustrated graphically some characteristics of the Binet test unappreciated up to that time. These characteristics included confirming a general spread of difficulty along the standard deviation based scale, but with major gaps (scale segments with no items) at the upper end of the scale. The scale highlighted a strong bunching up of items at about 2.5 units on the scale. The elegant ruler like presentation (Thurstone, 1925, p. 449) embodied the concepts needed to understand how items could be used as markers of learning development. The detachment of the scale from age cohorts to an absolute scale complements the exemplar scales of Thorndike. The graphic and the scaling process depend on the adequacy of the standard deviation unit as an appropriate unit.

Thurstone (1928) required that "the scale value of an item should be the same no matter which age group is used" (p. 119) and believed that Thorndike's approach was dependent upon the samples used, and thus not sample-distribution free. Engelhard reports that the Thorndike and Thurstone approaches yielded "essentially identical values when applied within one group" (Engelhard, 1984, p. 31). In multiple groups, Thurstone adjusted differences in the means of different groups. Thorndike assumed that the standard deviations in each of the ability distributions were equal (Engelhard, 1982; Holzinger, 1928). Using

Trabue's Completion Test data, Engelhard established that Rasch, Thurstone and Thorndike scaling processes produced linear scales of item difficulty that, within and across methods, were approximately invariant, and that each process had a linear relationship with the other two.

Plotting the data reported by Trabue and re-analysed by Engelhard, the relationship of Thorndike Units and logits is illustrated in Figure 2.2.

Figure 2.2 Trabue's Completion Test from Engelhard (1982)



Note: Data points are item difficulties for Trabue's Completion Test in Thorndike Units and as logits, based on Engelhard's (1982) conversion.

For Trabue's completion test data there is a strong linear relationship of Thorndike Units to logits. The scale analyses confirm that both Thorndike's and Thurstone's scale approaches are transformable to a logit scale. This confirmation implies that there is a consistency in their scaling approaches developed in the first quarter of the 20th century with those based on the more modern Rasch model. These early insights into scale concepts have yet to be fully realised in approaches to classroom assessment. However the progress map initiatives described in the subsequent chapters draw on this vision of scales, as do, less directly, the concepts of levels.

1916 A 'level' approach for composition

Hudelson (1916, p. 595) advocated the use of scales and urged that "we must start somewhere, and rather than go through all that has been done to work out established scales, we can use such standards as the Hillegas or the Harvard-Newton to fix our units of

measurement, much as the zero and boiling-points are established on a new thermometer by measuring it with an authentic one." (Hudelson, 1916, p. 595) He offers that

> for composition, by establishing one or two points we can, from them, derive the other degrees. These need not be fixed upon a percentage basis; in fact, my belief is that a standard should be made to run from 10 to 120 in point or degree of difficulty; and the ideal minimum for the first grade would then be 10, for the second 20, for the third 30, and so on, up to 120 for the twelfth year. With this as a basis, our problem would then become one of choosing typical models for each year. (Hudelson, 1916, p. 595)

This is an encapsulation of a vision that was taken up (or re-invented) much later as part of the curriculum and outcome descriptions encapsulated in the English and Australian curricula in levels models in the 1980s and 90s. In these models the assessment vision involved processes to help the teacher assess either where students are in their development (where they are located on the scales) or where an artefact (item) produced by the student is located. The Hudelson scale vision assumed equal linear increments of 10 units per grade. Assuming continuous improvement this would average about 1 unit per school month. The concept of an extended scale connecting the elements of the curriculum and scaling their difficulty, while not particular to Hudelson, anticipates later developments, particularly Wright's 'Academic Achievement Units', where, "say, 0 = entry into 1st Grade and 1000 = admission to College." (Wright, 2001, p. 784)

1950s - A new way forward.

In the period from the 1920s through to the 1950s the use of group and individual testing approaches increased. A range of journals was founded (*Psychometrika*, 1935; *Educational and Psychological Measurement*, 1941; *British Journal of Statistical Psychology*, 1947) (Du Bois, 1970) and a rich exchange of approaches to the development and analysis of test processes developed.

Somewhat isolated from this mainstream development, Georg Rasch, working mainly as a statistical consultant, was engaged to assist in the development of an intelligence test for the Danish military (Andersen & Olsen, 2000). This initial encounter with test and item difficulty led into a project on slow readers where children had been tested and remedially supported in their school years, and re-tested as adults in 1951. For various reasons (different reading tests, World War II) "it was not possible to evaluate the slow readers by standardisation as was the usual method of the time." (Andersen & Olsen, 2000, p. 10). Rasch needed to develop a method where an individual could be measured independently of which reading test had been originally used and in a way that could be connected to the 1951 test.

> The method was as follows: two of the tests which had been used to test the slow readers were given to a sample of schoolchildren in January 1952. Rasch graphically compared the number of misreadings in the two tests by plotting the number of misreadings in test 1 against the number of misreadings in test 2 for all persons … The graphical analysis showed that apart from random variations, the number of misreadings in the two tests were proportional for all persons. Furthermore this relationship held no matter which pair of reading tests he considered. (Andersen & Olsen, 2000, p. 10)

To account for the random variation Rasch developed a Poisson model. He was able to develop a model with two parameters, one for the subject (or person) and one for the item. Rasch had established that performance of students on a test could be related to the difficulty of the test and that it was "possible to deduce a distribution that depends only on the item parameters, but not on the person parameters" (Andersen & Olsen, 2000, p. 13). Through this process Rasch developed an approach to "bridge building", the process of placing persons on a common scale using different test instruments.

Subsequent consideration of similar problems led to the development of the Dichotomous Rasch model though the timescale for this development is not clear (Andersen & Olsen, 2000). The model, sometimes called the one-parameter logistic model (Allen & Yen, 1979), has become the basis for a family of models that maintain the independence of person from item.

This simple approach, when data fit the model, has opened up a range of possibilities that support the assertion that the placement of students on a (latent) scale can be estimated via a wide range of instruments or processes. The development of scales using the Rasch model makes feasible the use of teacher judgement as one of the 'instruments', as already described and advocated by others (Forster & Masters, 2004; Griffin, 2004). Engelhard (1984) established that the natural unit of the scale, the logit, already applies in the earlier scales developed in the period 1910 to 1930.

Summary

The early explorations of Fisher, Rice, Stone, Thorndike, Courtis, Hillegas, Trabue and Thurstone, as described above, set in train a technology and a concept of how learning in the classroom might be monitored in a consistent fashion. The approaches led to tools that teachers themselves could use in the classroom to monitor student development. Some of the developers were interested in the change of learning performance with grade and presented data that helped teachers and school administrators see assessment as an approach to observing and confirming individual student development over time. In many cases these data assumed latent scales of development, calibrated with equal interval units that bore (it can now be established) an approximate but direct relationship to the currently widely used

Rasch logit scale. For the useful observation of development over time, the learning unit must be of a consistent amount. That early developments in scaling bore a strong relationship to the now accepted log odds scale provides confidence that these early researchers were creating measurement scales with meaningful units.

This initial direct use of scaled techniques to assist holistic judgments by teachers and establishing estimates of scale positions of students on other developmental scales, seems to have lasted in this form until the late 1920s but was not well embedded in classroom practice. Assessment moved from a scaling tradition (Thorndike) to a test score tradition (Wood) in this period (Engelhard, 1992b). The production of an increasing range of paper and pencil tests was combined with the tendency to narrow quantification from measurement with extended rulers of learning, to become focused on pass-fail tests that established who crossed critical boundaries. This was achieved without optimising the meaning of the value of the underlying latent scale. The concept of a measurement scale was disregarded altogether by many teachers who instead perpetuated alpha grades, adjectives and/or percentages. With the development of approaches to assessment, particularly those initiated by Rasch in the 1950s and 60s, concepts of educational measurement were redeveloped. These led to the possibilities of extended scales of development and, in some implementations, the use of teacher judgment as the process for establishing the position of students on scales of development.

Although limited by manual data processing, and a reliance on paper records in the classroom, the early scale developers provided a process to support data informed pedagogy. The early beginnings of the conceptual steps in educational measurement, while concerned often with monitoring teachers more so than learning, have set a path that can be redirected back to helping teachers assess consistently and to record assessments in a systematic way. That the potential was not fully developed can be interpreted, by some, as a failure of the processes. For others, and in terms of this thesis particularly, the evidence is that appropriate thinking and concepts have existed for 100 years. Scaling and mapping learning, critical still in test technologies, has the potential to be applied to help teachers better understand and thus better observe and assess the learning of their students.

To summarise the perspective of the 1920s, McCall in his *How to Measure in Education* (McCall, 1922) argues that there are many reasons to see teachers' judgements (of the period) as inadequate or inaccurate in classifying students; this role of classifying now no longer a prime requirement of the classroom. However he comes to the view that teachers' judgements have importance.

> Teachers' marks are important because they are now and will continue for some time to be the most universal method of rating pupils. In fact, they may continue forever to

> be the criterion for classification [in the modern context read ‘assessment’], because teachers will soon be familiar with the simple mysteries of scientific measurement. They will themselves use tests with the same ease and fluency that they now use textbooks. More and more they will base their judgments upon objective rather than subjective measurement. When this time arrives teachers' marks will be not only as accurate as objective measurement, but they will be objective measurement plus something else. (McCall, 1922, p. 59)

To achieve this accuracy outcome requires support and encouragement for teachers, with the right frameworks and assessment tools. Curricula in levels combined with processes for assessing progress through each level, provide one model for doing this. The next chapter considers the more recent history of the development of levelled curricula, particularly in Australia in the 1980s and 90s that led to the teacher judgement data analysed in Chapters 7 and 8. A curriculum described in levels bears a conceptual relationship and a direct historical link (based on Hudelson) to the scale developments. A key attribute of early scales is the process to position individual and group summaries of learning status at scale values between the ‘prime’ scale markers. Initially, this attribute did not carry over to the Australian levels, in South Australia at least. A process to estimate progress from one level to the next is required to maximise the value of a level scale to teachers.