Int J Technol Des Educ (2012) 22:157–170
DOI 10.1007/s10798-011-9189-x

Comparative judgement for assessment

Alastair Pollitt

Published online: 14 December 2011

Abstract Historically speaking, students were judged long before they were marked. The tradition of marking, or scoring, pieces of work students offer for assessment is little more than two centuries old, and was introduced mainly to cope with specific problems arising from the growth in the numbers graduating from universities as the industrial revolution progressed. This paper describes the principles behind the method of Comparative Judgement, and in particular *Adaptive* Comparative Judgement, a technique borrowed from psychophysics which is able to generate extremely reliable results for educational assessment, and which is based on the kind of holistic evaluation that we assume was the basis for judgement in pre-marking days, and that the users of assessment results expect our assessment schemes to capture.

Keywords ACJ · Assessment · Judgement · Reliability · Adaptive Comparative Judgement

Introduction: a brief history of assessment

Judgement and marking: China[1]

From at least the Sui Dynasty (581–618), candidates for the imperial Civil Service were selected on merit using written examinations. At every level, from the lowest 'district' exams up to the highest 'palace' exams, the outcome was a grade for each candidate, together with a complete rank order amongst the highest grades, though the number of grades varied, from just two (pass/fail) at lower grades to five further up the system.

The most important principle was that the system was designed to be meritocratic, and extraordinary efforts and expense were put to ensuring that the best candidates in China

[1] This description is based mostly on the description of the late nineteenth century Qing system by Miyazaki (1963).

A. Pollitt (✉)
Cambridge Exam Research, Cambridge, UK
e-mail: alastair@camexam.co.uk

won through. The examination was, in theory if not always in practice, open to every (male) citizen, based solely on a specified set of 'classic' books. At every level, candidates' papers were anonymised, using seat numbers in place of names to avoid any risk of favouritism. From provincial level up papers were re-written by clerks, and checked by proof-readers, to remove the influence of calligraphic skills. Candidates—and officials—were body-searched on entry and isolated in the examination compound for up to 3 days at a time to minimise cheating: punishments as extreme as execution were applied to candidates and officials if any cheating was discovered.

Assessment was fundamentally the judgement of the quality of answers to open questions. At most levels, the ranking of top candidates was determined by the judgement of a single chief examiner who, at the highest level of the palace examination, was the emperor himself. At lower levels, the numbers of candidates involved meant that associate examiners contributed, usually by screening out the weaker candidates so that the chief only considered the best. Even so, the amount of work involved could be enormous, requiring eighteen associate examiners to evaluate as many as 20,000 candidates in the largest provincial exams. Arithmetic was used only in cases where a few grades from different papers had to be averaged. Multiple examining was used only at the highest level: in the palace exam, where there were fewer than four hundred candidates, about eight high officials read every paper and, by argument, agreed on the final ten. These were then presented to the emperor, who determined the final ranking.

Thus, in China, the two key problems for a meritocratic examination system—the risk of bias and the practical challenge of large numbers of candidates—were met with by applying huge bureaucratic resources. Thousands of officials, clerks and soldiers, including governors, ministers and the emperor himself, gave several days of their time to the process. Until the abolition of the system in 1905, as the modernisation movement caused China to look abroad for new models, one principle remained absolute—candidates should be *judged* on the quality of their work.

Judgement and marking: Europe

In medieval European universities assessment was generally oral, and the outcome for a student depended on the judgement of his tutors and other university examiners. In general, passing the examination was simply a licence to teach in a university, and the results had no particular importance for other careers. In the later decades of the eighteenth century, however, the same issues gradually became salient as in China. The response, however, was different, and a shift away from judgement began in Cambridge.

As the age of enlightenment developed into the industrial age the number of candidates grew, and the workload of examiners increased proportionally. Written examinations were introduced in Cambridge, gradually, in the second half of the 18th century, but were at first less important than oral interviews; they consisted principally of mathematics, with a small component of theology. By the end of the examination week, college and university tutors and officers had agreed on a rank order and classification of all of the candidates (Wordsworth 1877). Lacking the resources and the prestige the Chinese centralised system could provide, the university needed somehow to spread the load. Secondly, the controversy surrounding the award of first place to Massey, rather than Watson, in 1759 showed both the growing prestige associated with coming top in the examination and the risk of favouritism or bias in a system based on personal judgement and argument (Watson 1818; Haley and Wothers 2005).

The solution to both problems was the invention of *marking* (Haley and Wothers, ibid; Stray 2001). William Farish, who served as University Proctor—an officer responsible for many functions, including the conduct of examinations—in 1792 and 1793 (and later Professor of Chemistry and Engineering) required his examiners to assign specific marks to individual questions so that, by adding them up, a final rank order would be arrived at quickly and fairly.

Thus 'judgement' was replaced by 'marking' in Cambridge, and the practice quickly spread through the English-speaking world and then beyond including, in the twentieth century, to China. With the emergence of statistical analysis, measurement theory became entirely a theory of test scores, and the concept of reliability came to be the main concern of test developers. Where judgements survived—in the assessment of art, speaking or other performances—it was assumed that they must be expressed as scores, as if they had arisen from marking schemes. For 200 years or so 'true score theory' dominated educational assessment. Whether this was beneficial to education or not is a matter for debate.

Comparative judgement

Is it necessary to abandon professional judgement in order to achieve reliability in educational assessment? As will be explained, the germ of an alternative that avoids this problem has been available for 80 years, but without computer technology and, in particular, web-based services, the solution has been too limited to be of much practical importance.

Thurstone and the measurement of psychological phenomena

The method is based on papers published from 1925 to 1935 by the Swedish–American psychologist Louis Thurstone, which presented several possible ways to construct measurement scales for psychological, rather than physical, phenomena. In these early days of empirical psychology, a major research activity was the search for laws connecting the mental perception of 'greyness' or 'loudness' to their physical correlates of the percentage of black and white or the pressure of a sound wave, which could be measured using standard engineering techniques. In addition, Thurstone sought ways of measuring purely subjective properties such as the strength of individuals' opinions or feelings on various matters and, more pertinent to our present concern, the subjectively perceived quality of things "such as handwriting specimens, English compositions, sewing samples, Oriental rugs" (Thurstone 1927a, p. 376). If psychology was to become a true science, it was essential that it learned how to measure the things it studied.

Of all Thurstone's models, the one that has proved most useful is known as the "Law of Comparative Judgement" (Thurstone 1927b). In principle, he assumed that a person perceiving some phenomenon will assign to it some instantaneous 'value', a value which may vary randomly to some degree from occasion to occasion, and that when we ask them to chose the 'better' of two such phenomena it is these values that are compared.

The original article describes in detail how Thurstone derived the model, and how it may be simplified, making several plausible assumptions, into a more practical form which he called 'Case 5'. This was used widely in psychological research, and in a few applied research contexts, in the following decades. Its usefulness was limited, however, because Thurstone assumed (like all psychologists and statisticians of his time) that the essential randomness in his data should be modelled using a *normal* distribution; the normal distribution is mathematically difficult to use computationally. In the 1950s the Danish educational statistician Georg Rasch developed a series of mathematical models for

educational tests (Rasch 1960/1980) using other distributions, and when Andrich (1978) demonstrated that Rasch's *logistic* model was equivalent to Thurstone's model it became possible to write relatively simple computer programs to apply the method of comparative judgement. It is this logistic model that we now use, and refer to as Thurstone's CJ model.

The logistic model for Comparative Judgement

In its modern logistic form, Thurstone's Case 5 model can be written:

$$\log \text{odds}(\text{A beats B}|v_a, v_b) = v_a - v_b \tag{1}$$

On the right hand side are v_a and v_b, which represent the 'value' of two objects, A and B. Thus, for example, the left hand side shows that, if we knew how 'nice' the oriental rugs A and B really were, then we could calculate the probability that, in any single comparison, a judge will decide that A is 'nicer than' B. More precisely, the difference $v_a - v_b$ is the natural logarithm of the odds that rug A will 'win'. The anti-log of the difference will then give us the odds that rug A will win, and we can convert from odds to probability in the usual way. The result is the probability form of the model:

$$\text{prob}(\text{A beats B}|v_a, v_b) = \frac{\exp(v_a - v_b)}{1 + \exp(v_a - v_b)} \tag{2}$$

In practice, of course, the reasoning goes the other way. The left hand side represents *data*, as it is the decisions of the judges that indicate the probability of a win for rug A or rug B—from these data we can then estimate the relative values of the various rugs. There are several important points to make about this model.

1. The values v_a and v_b are relative values, since only the differences between pairs of them are obtained from the data. We therefore need to fix an arbitrary point—like zero on the Celsius scale, which was arbitrarily chosen to represent the freezing temperature of pure water—before we can attach numbers to the scale. Similarly the unit of the scale, although it is formally defined by the probability relationships above, is essentially arbitrary and, like the units of Celsius and Fahrenheit, can be changed to any convenient size.
2. When Thurstone originally used this method, he took care to collect several, or many, comparisons of every possible pair of objects he wanted to scale. This allowed every probability in Eq. 2 to be estimated reasonable accurately, before they were all combined to estimate the final values of all the objects. Over the years we have found that the system is extraordinarily robust; it is not necessary even to observe one judgement of every possible comparison. If every object is compared about ten times to suitable other objects, this will generate a data set that is adequate to estimate the values of every object on a single scale.
3. This in turn shows that Thurstone's original psychological underpinning for the method is unnecessary—whatever psychological reality underlies comparative judgement is general enough to apply in many contexts that do not meet the conditions that Thurstone assumed for his research.
4. The 'objects' compared can, in principle, be of almost any kind. It seems from experience, and the robustness of CJ data, that the key requirement is not that there be an instantaneous 'response' of the kind Thurstone imagined but only that judges be able to form a holistic evaluative judgement of something against a notional scale that is a shared consensus of all of the judges. We can imagine this happening in many educational contexts, including the

evaluation of performances in writing, speaking, art or dance, or of portfolios in design, journalism or art, or of project reports in science, geography or technology. Whether the necessary consensus exists is, in fact, an empirical question in each case.

5. The values for the objects can be estimated using an iterative maximum likelihood (ML) procedure. The commonest way to analyse CJ data today is to use *FACETS* (Linacre 1994), but this program requires all of the data to be collected before analysis begins. The work reported in this issue uses a variant of CJ called *Adaptive* Comparative Judgement, for which a more robust procedure was needed. The principles and equations that underlie the ML procedure used are described in Wright and Masters (1982), and the algorithms were written and implemented by the author in collaboration with TAG Developments (see below). The application of CJ for assessment and learning depends fundamentally on the power given by this adaptive model, which is described below.
6. The model is conceptually very simple, and strong in the statistical sense that only one parameter is needed for each object to fully describe the data. Strong models imply powerful statistical control, since any deviation from prediction can be detected and evaluated. The theoretical assumptions Thurstone made to derive his Case 5 are testable: specifically it is assumed that judges are equally good in their ability to discriminate between objects, that every object is equally 'discriminable', and that each judgement is independent of every other one. Each of these can be checked through analysis for *judge misfit*, *object misfit*, and *bias* respectively. This will be discussed in more detail below. Given the context of this issue, for the rest of the paper I will generally refer to 'objects' as portfolios.

Early uses of comparative judgement in education

Research

The first published use of CJ in education was in a study into the nature of spoken language proficiency (Pollitt and Murray 1993). Video recordings of the test performances of five students were sliced and spliced to give pairs of 2 min segments, which were judged by a team of six judges. A qualitative study was included, using a clinical construct-elicitation procedure based on Kelly's Personal Construct Theory (Kelly 1955), to determine what actually defined the construct 'speaking proficiency'—as experienced by assessors while they were engaged in the process of judgement rather than that of teachers or academics reflecting or rationalising from a detached perspective. The decisions were used to construct the scale, and the interview evidence was then used to interpret and illustrate it.

Comparability studies

In a 1995 research seminar of the UK examination boards several speakers expressed concerns with their 'cross-moderation' methodology for comparing standards across boards in what were generally supposed to be equivalent examinations. Adams (1995), for example, declared: "The data are rough and ready. To use elegant and sophisticated techniques on such crude data may be at best unaesthetic and at worst daft". Pollitt introduced the CJ method to them as an alternative, heralding what Adams (2007) refers to as "the paired comparison revolution". The first study using it was reported in D'Arcy

(1997). Since then, CJ has been the only empirical method used in the UK for assessing comparability (see Bramley 2007, for a detailed description of this work).

Adaptive comparative judgement (ACJ)

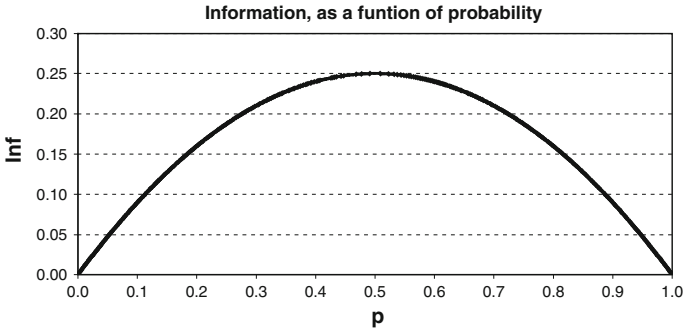
The idea of computer-based adaptive testing with multiple-choice items is now fairly familiar (Weiss and Kingsbury 1984; Wainer 2000). After each item that a student responds to, an algorithm calculates an updated estimate of their ability; they are then presented with a new item whose difficulty is similar to that ability estimate. The procedure continues, with a new estimate and a new 'optimal' next item, until some criterion determines that the measurement is finished. In many cases, the criterion relates to accuracy: stopping the process when the remaining error or uncertainty is small enough, or when the confidence that a candidate has passed or failed is high enough. The principal benefit of adaptiveness in testing is efficiency, since the required level of accuracy is reached after only 40–50% as many items as with traditional standard tests, though concerns about covering the specified content mean that adaptive tests in education are often longer than is strictly necessary.

The same advantage, but without the concern, applies to ACJ. Comparing two scripts, two portfolios, or two performances, we will learn more about them if they are reasonably similar in quality than we will if they differ greatly. A useful metaphor to consider is sports competition. In general, it is more interesting to watch two evenly matched football teams compete, since the result is more unpredictable, than to watch one from a higher league against one from a lower one; the result of a close match tells us more than we knew before about the relative quality of the two teams, while a higher league team beating a lower one merely confirms our pre-judgement. The analogy should not be pushed too far, though, as sports often have unpredictably designed into them, in a way that is not normally intended in educational assessment.

The increase in efficiency is described by the statistical concept of *information*. For our case, the information in a result, I, is defined by

$$I = p(1 - p) \tag{3}$$

The result is either a win for the first portfolio and a loss for the other, or vice versa. In either case, the information is the product of the probability of one outcome and the probability of the other. Since we do not allow ties, if the probability of one is p then the probability of the other is $(1 - p)$. This function is maximised when $p = 0.5$, but remains reasonably high for some distance to either side of this, as the graph shows. We conclude that choosing portfolios for comparison so that the probability of one of them winning is anywhere between 0.3 and 0.7 will give us between 0.21 and 0.25 units of information, or at least 84% optimal efficiency .



Experience has shown that true optimisation demands a balance between two conflicting principles. Information is maximised when $p = 0.5$, but this makes the decision process difficult for judges: their decision will be easiest when the two portfolios are very different in quality and p is close to 0 or 1. We have chosen to optimise the assessment process by choosing a comparator for each portfolio so that p is approximately 0.67 or 0.33.

A web-based system has been developed for the e-scape project in association with TAG Developments, a division of BLi Education, to facilitate the use of ACJ. It is designed to present judges with pairs of portfolios chosen to meet this optimisation criterion. To run it, we needed to solve the pragmatic problem of finding 'starting values' for the portfolios.

Before any judgements have been made, of course, nothing is known about the quality of each portfolio and so the initial pairings must be purely random. As soon as every portfolio has been judged once the system can do better than random, by presenting pairs which have either both won or both lost their first comparison, since these are likely to be closer in quality than a random pair would be. This strategy can be repeated for several rounds, choosing pairs with the same, or a similar, number of wins so far, and each time it will be a little more efficient as the better and poorer portfolios separate out. Eventually, however, enough data will have been collected to allow a computer to estimate each portfolio's value using the Rasch model in Eq. 2; then the efficiency rises dramatically as the values are known with increasing accuracy and the pairing is closer to optimal.

The 'number of wins' method for the early rounds was borrowed from competitive chess, where it is known as the 'Swiss tournament' system, and is designed to ensure that most players, most of the time, meet players of a similar standard. At first we ran six 'Swiss' rounds, in order to give the program enough data to work on, but improvements to the algorithm have made it possible to estimate the Rasch values for the portfolios after only three or four. This brings a further improvement since the system is able to 'chain' comparisons once the Rasch method is in use. Suppose a judge has compared portfolios P77 and P93: the next decision they are asked to make may be between P93 and P21, and then P21 and P67—each comparison only involves one new portfolio rather than two. This may make little difference when comparing oriental rugs, but the time saved when comparing complex portfolios is substantial.

There is inevitably a concern that this may contradict the assumption that every judgement is independent of all others; making two comparisons with the same portfolio raises the possibility that the first decision may influence the second in some way. No one has yet found convincing evidence of a bias of this kind and, given that it would apply in a small way randomly to every portfolio, the gain in efficiency seems well worth the risk. Nevertheless, we will continue to look for any evidence of chaining bias.

As mentioned earlier, the ACJ system has proved remarkably robust. In Thurstone's original work, every object was paired for comparison with every other one, and each pairing was judged by several judges. Thus the 2008 Phase 3 e-scape trial (Kimbell et al. 2009), involving 352 portfolios, would have meant

$$(352 \times 351 \div 2) = 61,766$$

different pairings, and about 618,000 judgements in total would have been needed. In contrast, in Adaptive CJ assessment each portfolio is compared to only a few others considered close to it in quality. In the e-scape trial with 352 portfolios, only 3,097 comparisons were made, with each portfolio compared to only about 17 of the other 351, and most of these comparisons were made only by one judge. Nevertheless, a very

consistent set of measures was obtained, with an internal reliability of 0.95. This demonstrates the extraordinary power of the adaptive procedure.

Statistical quality control

Analysis of residuals

When a judge compares two portfolios, Eq. 2 can be used to measure the degree of 'surprise' in their decision. It gives us the 'predicted' outcome, $p_{a,b}$, where 'expected' has the standard statistical meaning of the average outcome if we could repeat the judgement many times. The value of $p_{a,b}$ is calculated from the final estimates of the quality of portfolios A and B, and will be a real number between 0 and 1. However, the 'observed' outcome will be *either* 1, if A wins, or 0, if B wins. The difference between the observed outcome and the expected one is called the *residual*, and this is at the heart of all the analyses that can be carried out to monitor measurement quality.

$$\text{Residual}_{j,a,b}: \quad X_{j,a,b} - p_{a,b} \;|X = 1, 0 \tag{4}$$

This is the residual when judge j compares portfolios a and b. The residual is first standardised by dividing by the square root of the Information in the judgement (from Eq. 3):

$$\text{StdRes}: \quad z_{j,a,b} = \frac{\text{Residual}_{j,a,b}}{\sqrt{p_{a,b}(1 - p_{a,b})}} \tag{5}$$

These standardised residuals can now be aggregated in any appropriate way and interpreted as a mean Chi-square statistic.

Misfit

For example, the performance of any judge can be evaluated by averaging the squared standardised residuals from all the judgements they made. Whenever a judge decides in favour of the portfolio that the consensus says is 'better', the residual will be smaller than 0.5; when they favour the 'poorer' one the residual will be larger. If all the residuals for a particular judge are summarised, the size of the resulting statistic indicates the degree to which that judge tends to deviate from the consensus, or *mis-fits*. The residuals can be summarised in two ways, producing what are known in the Rasch literature as *Infit* and *Outfit Mean Squares*. The Infit Mean Square is calculated as follows.

First, each standard residual is squared and then weighted by its Information:

$$\text{WtdSqRes}_{j,a,b}: \quad wz^2_{j,a,b} = I_{a,b} \times z^2_{a,b} \tag{6}$$

These are summed across all the judgements made by judge j, and divided by the sum of the information, to give the Infit Mean Square:

$$\text{InfitMS}: \quad \text{wms}_j = \frac{\sum_{\text{judge}=j} wz^2_{j,a,b}}{\sum_{\text{judge}=j} I} \tag{7}$$

This statistic is reported in the TAG e-scape system (under *full report*) as:

Number	Name	Count	MnRes	UnWMS	UnWZ	WMS	WZ
041 A.	Judge	156	0.48	0.97	0.37	0.99	-0.47
042 A.	Other	179	0.48	1.62	0.89	**1.13**	3.74
Mean:		0.47	1.87	1.42	**1.02**	0.53	1.87
SD:		0.01	3.27	3.03	**0.06**	1.47	3.27

The Infit mean square (WMS) is interpreted as a mean Chi-square, and a conventional criterion is to treat as mis-fitting any judge whose value exceeds the mean plus two standard deviations: in this case the criterion would be $1.02 + 2 \times 0.06 = 1.14$. Thus '`Judge A. Other`', with a value of 1.13, would be considered close to the borderline for conforming to the consensus, but just acceptable. When a judge's WMS exceeds the limit it may be worth exploring how they made their judgements (they may see some value that all the others miss, or balance originality versus practicality differently) but, in the end, reliable assessment requires that the judges are consistent with each other.

Because adaptive testing generates residuals that are not randomly distributed, the statistics must be treated with some caution. At present, the weighted mean square statistic seems to be the most trustworthy, and is recommended. Work in progress will make it easier for users to interpret these statistics.

A similar procedure gives a 'misfit statistic' for each portfolio, by summarising the weighted squared residuals and Information over all the comparisons that it is involved in. If a portfolio's WMS exceeds the criterion of mean plus two standard deviations, this means the judges did not judge it consistently, some considering it 'better' than others did. Significant portfolio misfit indicates a specific difference between judges in how they would rank order the portfolios, arising from a difference in how they understand or value the trait they are trying to assess.

There will normally be fewer comparisons for each portfolio than comparisons by each judge, so the statistic is rather less sensitive, but it is still good enough to identify any problem portfolio that will need special attention. There may be some unusual feature in it that is valued more highly by some judges than others, or some missing evidence that only some judges are willing to assume the student knew. In such a case, and particularly if the portfolio lies close to a grade boundary, the statistic suggests that more judgements are needed, perhaps from particular, senior, judges who take the responsibility to rule on difficult cases. And, again, there is an opportunity here for some research to understand better how students and judges understand the trait being assessed.

Bias

These two misfit statistics are special cases of a more general form of analysis. Because every judgement is assumed to be strictly independent, the summation can be carried out over *any* subset of comparisons, not only those involving one judge or one portfolio. If there is any reason to suspect that a subset might be systematically different from the rest, the weighted squared residuals for that subset can be summarised in the same way, giving a significance test for bias.

Examples of the kind of hypothesis that might be explored include that male judges favour male students' work, or that judges with a post-graduate qualification give more credit to creative solutions than other judges, or that teachers favour portfolios from their

own school. Pollitt and Elliott (2003) reported a study of comparability data investigating the hypothesis that a judge will favour scripts from their own exam board when comparing them to scripts from other boards: in that study just one out of nine judges showed a statistically significant but substantively small amount of 'home' bias.

Advantages of comparative judgement in educational assessment

The method has several advantages over marking methods. In this section I will describe only the technical bases for some of these.

Validity

Assessors cannot afford to ignore validity, or to leave it to others to validate their tests.

> Validity … is built into the test from the outset rather than being limited to the last stages of test development, as in traditional criterion-related validation. (Anastasi 1986: p3)

Valid assessment requires the design of good test activities and good evaluation of students' responses using good criteria. The traditional concept of construct validity requires that we are clear about what we want to measure, that the teachers and students also are clear what that is, and that our assessment does indeed reward students for showing evidence of what we wanted to measure and them to develop or learn.

Since CJ requires holistic assessment, a very general set of criteria is needed—although they may include the ability to apply the general skills to a particular task. In much of our work we have relied on the 'Importance Statements' published on England's National Curriculum web site, such as:

The importance of design and technology

> In design and technology pupils combine practical and technological skills with creative thinking to design and make products and systems that meet human needs. They learn to use current technologies and consider the impact of future technological developments. They learn to think creatively and intervene to improve the quality of life, solving problems as individuals and members of a team.
> Working in stimulating contexts that provide a range of opportunities and draw on the local ethos, community and wider world, pupils identify needs and opportunities. They respond with ideas, products and systems, challenging expectations where appropriate. They combine practical and intellectual skills with an understanding of aesthetic, technical, cultural, health, social, emotional, economic, industrial and environmental issues. As they do so, they evaluate present and past design and technology, and its uses and effects. Through design and technology pupils develop confidence in using practical skills and become discriminating users of products. They apply their creative thinking and learn to innovate. (QCDA 2011)

These 160 words were designed to guide teachers, curriculum designers, assessors *and* students towards a common understanding of what is important in Design and Technology. It follows that using them as the essential criteria for assessing students' work should lead

to more valid assessment than any other criteria. ACJ judges are expected to keep their minds on what it means to be good at Design and Technology: there is no mark scheme to get in the way.

In Thurstone's original use of the method in psychophysics, the 'quality' being evaluated was quite simple—the perceived pitch of a musical tone, or the intensity of a negative feeling. He later applied it to measuring attitudes and values, which cannot have a simple physical substrate but represent the unconscious sum of several elements. In assessment we have shown that this can be successfully extended to assess the 'overall quality' of portfolios, project reports, or other complex objects. With these the 'quality' being measured is far from simple, and it is probably very difficult to describe it in anything other than quite general terms. The method demands and checks that a sufficient consensus exists amongst the pool of judges involved.

As mentioned earlier, a qualitative study using Kelly's personal construct elicitation method can be combined with ACJ to investigate the nature of this consensus (Pollitt and Murray 1993).

Reliability

The most obvious technical advantage of ACJ is the extremely high reliability that it can deliver in any context in which subjective judgement is appropriate. In ACJ analysis, reliability is calculated in the following way. For every portfolio, i, the analysis generates a parameter, v_i, and a standard error, e_i. From these, the standard deviation of the parameters is calculated, sd_v, and the root mean square average of the standard errors, *rmse*. We then define a separation coefficient as the ratio of the sd to the *rmse*:

$$\text{Sep Coeff}: \quad G = \frac{\mathrm{sd}_v}{rmse} \tag{8}$$

If we think of the *rmse* as the amount of 'fuzziness' in the scale, then G tells us how the spread of the measures compares to this—the higher the value of G, the more clearly separated the objects are. G can be directly converted into an analogy of the traditional reliability index:

$$\text{Reliability}: \quad \alpha = \frac{G^2}{(1+G^2)} \tag{9}$$

Reliability refers to the precision of measurement, and is normally reported in one of two forms—*relative* or *absolute*—a distinction that comes from generalizability theory (for a full presentation of this theory, see Cronbach et al. 1972, or Shavelson and Webb 2000). The reliability coefficient in Eq 9 is equivalent to Cronbach's *alpha coefficient* (Wright and Masters 1982, pp. 134–136), which is usually considered to be relative.

In assessment by marking, where generalisability theory was developed, relative reliability describes how well the ranking generated by a single typical marker agrees with that of the consensus of markers, rather than how absolutely accurate a single score is. For example, any two reasonable thermometers will give more or less perfect relative reliability in a set of temperature measurements. Absolute reliability measures the extent to which the numerical outcomes of different measurements are equivalent—if one of the thermometers uses the Celsius scale and the other the Fahrenheit scale their *absolute reliability* will be quite poor.

However, in ACJ the measures are reported on a common scale which is the consensus from all the judges, and so much of the difference between relative and absolute reliability disappears: in effect, the reliability in ACJ corresponds to what would be achieved if each portfolio were marked by many markers rather than just one or a few. In the limit, if *every* teacher in the population of teachers takes an equal part in setting the consensus, the ACJ coefficient will be completely absolute. In practice, since all generalisability studies also involve only a sample of markers, ACJ reliability is as absolute a measure as any called 'absolute' in generalisability studies.

The clearest demonstration we have so far of the superior reliability of ACJ comes from two large studies of English pupils' writing. In one study—using marking carried out in Australia—Baker et al. (2008) reported an average absolute reliability of 0.85, and an average relative reliability of 0.93 (these figures are rather higher than studies within the UK for the same tests have typically obtained). In our (not yet published) study using ACJ through the TAG/e-scape 'pairs engine', albeit on a rather different sample of pupils and writing, the absolute reliability figure was 0.96. Even if it were considered as relative reliability this value would be very high, but as an absolute measure it is extremely so.

The comparison of these two studies makes two important points. First, the ACJ system is capable of generating extremely reliable scores, far higher than traditional marking can. Secondly the data are very simple, just the comparative judgements of experienced teachers who were focused wholly on making valid judgements, and who had approximately 2 h of training in the process—in contrast, the Australian markers averaged more than 15 years of experience in marking. The reliability of ACJ is a consequence of the constant focus on validity.

One very useful difference between ACJ and marking, with regard to reliability, derives from the principle that every judgement made is independent. Because of this, more data can be collected at any time, if needed to increase the reliability. The method can, within reason, reach any required level of reliability simply be asking more judges to make more judgements. But rather than simply increasing the overall reliability, these extra judgements can more usefully be targeted on particular portfolios, such as those that are close to a grade boundary; this would improve the accuracy of classification, which is what matters most in summative assessment, more than a general reliability coefficient. In most contexts we have seen so far, ACJ will deliver reliability well above that of marking for a comparable cost.

'Cancelling out' the judge

It is conventional, in studies of the assessment of Writing, to say that markers can go wrong in three ways, all of which apply more generally to the marking of any complex response. First, a marker may measure to a different standard, and mark more severely or more generously than the others. Markers need a considerable amount of training to instil the proper standard, and quite regular re-calibration to maintain it, since markers' standards are known to drift over time. Second, a marker may award the same average mark, yet discriminate more finely amongst the objects, in effect being more generous to the better ones and more severe to the poorer—or vice versa. Third, a marker may value different aspects of the overall quality differently, and so sort the students into a different rank order.

With CJ the first two of these do not matter. A judge whose overall standard is higher, or whose evaluation is finer grained, will still choose the same one as 'better' because they use the same standard and the same discrimination for both. In effect, the characteristics of the judge are cancelled out by the design of the data collection (Andrich 1978). It is this

cancellation, together with the involvement of multiple judges for each portfolio, that makes the reliability coefficient absolute rather than relative.

Only the third kind of error remains, since it is possible that different judges will conceive the 'overall quality' differently. Since there is no need to worry about standards or discrimination, any examiner training, and the investigation of examiner misfit, will be entirely focused on this issue, which is fundamentally about validity. The focus on validity also gives ACJ great potential for promoting learning through peer assessment, where the 'examiners' making the judgements are the students themselves. Then, any student whose decisions are not consistent with the others does not share their understanding of what is good quality work. Discussion and debate amongst the students and teachers about particular cases can be a powerful way to develop a professional understanding of standards in students.

Future improvements

We are exploring ways to replace the Swiss-system with other starting values. Possibilities include estimated grades provided by teachers, which are standard in the UK school examination system, or any prior performance indicators that may be available. An interesting idea is to ask the students themselves to provide initial estimates, since there is no advantage to them in over- or under-estimating their standard. Any information, even if not very reliable, is likely to be better than starting randomly.

A further strategy we have not yet implemented is referred to as 'focusing'. For this we need grade boundaries or pass scores to be set: then the system can stop asking for judgements of any portfolios that are clearly in one category, and instead focus on those close to a boundary. It is possible, in some contexts at least, to transfer a boundary standard from an old test to a new one by including some objects from the old test for comparison with the new ones.

Conclusion

Comparative Judgement is well established as a measurement technique, using standard Rasch software for analysis. Adaptive Comparative Judgement is new, and its potential for educational assessment is only beginning to be realised. It is clear, however, that it is capable of providing assessment of complex portfolios or performances with a level of reliability that has never been achieved through marking. For this kind of assessment, marking was never appropriate, and ACJ allows us to return to using proper human professional judgement to capture the essential quality in a student's work. What could be fairer than that?

References

Adams, R. M. (1995). *Analysing the results of cross-moderation studies.* Paper presented at a seminar on comparability, held jointly by the SRAC of the GCE boards and the IGRC of the GCSE groups, London, October.

Adams, R. (2007). Cross-moderation methods. In P. Newton, J. Baird, H. Patrick, H. Goldstein, P. Timms, & A. Wood (Eds.), *Techniques for monitoring the comparability of examination standards.* London,

QCA. Available (26/09/2011) at: http://www.ofqual.gov.uk/files/2007-comparability-exam-standards-h-chapter6.pdf.
Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.
Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*, 451–462.
Baker, E. L., Ayers, P., O'Neill, H. F., Choi, K., Sawyer, W., Sylvester., R. M., & Carroll, B. (2008). *KS3 English test marker study in Australia*. Final report to the National Assessment Agency of England, London, QCA.
Bramley, T., (2007). Paired comparison methods. In P. Newton, J. Baird, H. Patrick, H. Goldstein, P. Timms, & A. Wood (Eds). *Techniques for monitoring the comparability of examination standards*. London, QCA. Available (26/09/2011) at: http://www.ofqual.gov.uk/files/2007-comparability-exam-standards-i-chapter7.pdf.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley Lawrence Erlbaum Associates.
D'Arcy, J. (Ed.). (1997). *Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.
Haley, C., & Wothers, P. (2005). In M. D. Archer, & C. D. Haley (Eds.), *The 1702 chair of chemistry at Cambridge*. Cambridge: CUP.
Kelly, G. A. (1955). *The psychology of personal constructs* (Vol. I and II). New York: Norton.
Kimbell, R., Wheeler, T., Stables, K., Sheppard, T., Martin, F., Davies, D., Pollitt, A., & Whitehouse, G. (2009). *e-scape portfolio assessment: phase 3 report*. London: Technology Education Research Unit, Goldsmiths, UL. http://www.gold.ac.uk/teru/projectinfo/projecttitle,5882,en.php.
Linacre, J. M. (1994). *Many-facet Rasch measurement*, 2nd ed. Chicago: MESA Press.http://www.rasch.org/books.htm.
Miyazaki, I. (1963). *China's examination hell: the civil service examinations of imperial China* (C. Schirokauer (1976), Trans). New York: Weatherhill.
Pollitt, A., & Elliott, G. (2003). *Monitoring and investigating comparability: A proper role for human judgement*. Invited paper, QCA comparability seminar, Newport Pagnall. Qualifications and curriculum authority, London. Available at: http://www.camexam.co.uk/.
Pollitt, A., & Murray, N. L. (1993). What raters really pay attention to language testing research colloquium, Cambridge (Reprinted from M. Milanovic, & N. Saville (Eds.), 1996, *Studies in language testing 3: Performance testing, cognition and assessment*. Cambridge: Cambridge University Press).
QCDA. (2011). *Importance of design and technology key stage 3*. http://www.education.gov.uk/schools/teachingandlearning/curriculum/secondary/b00199489/dt/programme. Accessed: 9 Dec 2011.
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. Reprinted as 2nd ed., 1980, Chicago: University of Chicago Press.
Shavelson, R., & Webb, N. (2000). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research, Chapter 18*. London: Lawrence Erlbaum Associates.
Stray, C. (2001). The shift from oral to written examinations: Cambridge and Oxford 1700–1900. *Assessment in Education, 8*, 33–50.
Thurstone, L. L. (1927a). Psychophysical analysis. *The American Journal of Psychology, 38*, 368–389.
Thurstone, L. L. (1927b). A law of comparative judgment. *Psychological Review*, *34*, 273–286 (Reprinted as Chapter 3 from Thurstone, L. L. (1959). *The measurement of values*. Chicago, IL: University of Chicago Press).
Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
Watson, R. (1818). *Anecdotes of the life of Richard Watson … written by himself at different intervals, and revised in 1814*. Published by his son, Richard Watson, L. L. B., prebendary of Landaff and Wells. London: T. Cadell and W. Davies.
Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.
Wordsworth, C. (1877). *Scholae academicae*. London: Frank Cass.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press. http://www.rasch.org/books.htm.