Monitoring and investigating comparability: a proper role for human judgement.

Alastair Pollitt

Gill Elliott

Research and Evaluation Division

University of Cambridge Local Examinations Syndicate

1 Hills Road

Cambridge

Abstract

In this paper we will discuss both the currently-used 'Thurstone' methodology for establishing comparability and the previously used 'home and away' ratification method. A brief description and history of the use of human judgement in comparability studies using both these methods will be accompanied by a detailed exploration of issues arising from them.

We propose to present the findings of an initial investigation into the extent and predictability of 'home Board' bias amongst judges, derived from a recent inter-Board study.

In conclusion, we shall consider what the ideal role is for human judgement in comparability, and also whether human judgement should be utilised at all during Awarding meetings.

Disclaimer

The opinions expressed in this paper are those of the author and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate or any of its subsidiaries.

Note

Part of this paper uses data collected by the Assessment & Qualifications Alliance on behalf of the Joint Council for General Qualifications. We are grateful for their permission to use the data.

Contact details

Alastair Pollitt & Gill Elliott, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU.
pollitt.a@ucles.org.uk and elliott.g@ucles.org.uk.

Monitoring and investigating comparability: a proper role for human judgement.

Alastair Pollitt
Gill Elliott

Research and Evaluation Division
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge

Background Paper

This background paper is in two parts: Part I: An introduction to the methodology provides an introduction to the use and history of cross-moderation techniques of comparability. A later draft of some of this material has been published by Elliott and Greatorex (2002) as: "A Fair Comparison? The Evolution of Methods of Comparability in National Assessment". The reference for this and other salient papers are provided for those participants not already familiar with the methodology in question. We do not intend to revisit this material at this level of detail during our presentation, which will concentrate upon the issues raised in part II: A proper role for human judgement.

Part I: An introduction to the methodology

Introduction

Comparability research is concerned with the monitoring and maintaining of parallel forms of qualification, which occur:

- over time
- between subject areas
- between awarding bodies
- between syllabus specifications
- between units

- between modes of examination (e.g. computer based testing versus pen and paper, coursework versus practical options.)

The use of cross moderation in comparability studies by Awarding Bodies has been common practice over the past quarter-century since the first study using this technique was undertaken around 1975. Cross-moderation refers to studies where experts from different qualifications make judgements about the relative standards achieved by candidates in different assessments. The purpose of such studies is both to monitor assessments which are expected to be comparable (and thus facilitate adjustments to keep the assessments in line), and to provide public evidence of the extent to which comparability is achieved.

Inevitably the methodology used has changed over the years, as new techniques of analysis are brought into play. In this paper we report on both 'home and away' ratification methodology and Thurstone pairs.

Ratification methodology (using a personal standard as a yardstick).

This methodology was used initially by Houston (1975) and has since been used in many internal studies within UCLES, as well as in most of the inter-board studies until recently. It involves the use of teams of judges, each team representing one of the examinations under scrutiny. The examiners are asked to fix in their minds the 'standard' that they would expect a typical candidate to achieve at the given boundary in the examination that they represent. They then make a series of judgements as to whether that script is at the boundary as defined by their own individual internalised perspective of standards, or whether it is above or below that boundary.

It is then assumed that examiners will 'favour' (i.e. judge to be most stringent) the standards of their own examination (Massey & Newbould, 1977) producing a 'home' advantage. Where the examinations under scrutiny are each judged to be more stringent by the home examiners a draw may be declared and the examinations considered comparable. Where both 'home' examiners and 'away' examiners agree that one particular examination is more or less stringent than another, the examinations are judged not to be comparable.

Balanced numbers of examiners from each examination under scrutiny are necessary, and it is also important to use examiners who are attached to only one of the examinations. It is difficult to use examiners who work on both examinations under scrutiny, because they do not have a single clear internalised standard upon which to base their judgements. It is also not possible to use independent scrutineers, because they do not have the necessary internalised standards of a particular examination.

Results from this methodology are translated into scores (often 1, 0 and –1, but other variants have been used) and scores presented in tabular form, with row and column totals. Significant differences between standards are established by the use of non-parametric statistics such as Kendall's co-efficient of concordance (*W)*.

Thurstone Pairs Methodology

The Thurstone Pairs methodology is in many respects similar to the ratification method, in that a number of examiner judges are recruited to carry out the research exercise at

similar boundary points. However, in this case examiners compare one script directly against another and judge which is the better. No ties are allowed. By combining all the judgements, a Rasch model can produce statistics showing the relative position of every script with every other, in the judges' estimation. At the top are the scripts consistently judged to be of highest standard and at the bottom those judged weakest. The scripts would appear clumped together on the scale if all the scripts were judged to be very similar. The model also shows the relative positions (to each other) of the judges, and identifies particular scripts (and particular judges) where conflict arose, or emerging trends were not maintained (misfit).

Practicalities of comparability studies.

Recruiting examiners.

It is the expertise of the judges which gives this form of comparability study some of its validity. Human judgement can in this instance enable the examiners to make judgements about the overall quality of work as demonstrated on a particular pair of scripts, whilst at the same time allowing for differences between the stimulus material (questions) and the curriculum generally. Some comparability studies have included independent judges as well as judges from the Awarding Bodies. Forster and Gray (2000) found that there were no statistical differences between the judgements made by the independent and Awarding Body judges.

The number of examiners selected varies from exercise to exercise. As always, the more judges there are the more confident we can be about the result. However, for practical and financial reasons it normally falls at between 3 and 6 examiners per examination under scrutiny. There is little evidence as yet about what constitutes an ideal number of examiners/scripts.

Selecting candidates' work.

Scripts selected for comparability studies are usually on or near the pertinent boundaries (i.e. those boundaries which examiners are accustomed to judging at awarding). More recently it has also been considered advantageous to incorporate scripts from a known range of marks around the boundary. This means that if standards are not found comparable, the issue of 'by how much' may be addressed. For an example of this practice see Bell et al. (1997).

It is possible to compare assessments in two ways – either 'whole candidate's' scripts, or separately by paper/unit. 'Whole candidate' scripts involve selecting the entire work of a given candidate which contributes to the overall assessment – thus all written papers or modules, oral or practical components must be judged by the examiners and their judgement made upon the holistic package. This can present both organisational and judgemental difficulties, for a number of reasons:

- There are many routes to achieve the overall borderline mark – which do we choose?
- For many oral/coursework options only a proportion of work is taped/held by the Awarding Body. This limits (and biases) the candidates we can select from, or necessitates requesting work from centres directly.

- It affects how examiners structure their approach to judging, given they may have to read, listen to tapes, or examine practical work.

If a perfectly balanced profile of work is desired (or if it is impossible to achieve the borderlines required using whole candidates), pseudo-candidates may be used – a 'whole candidate' made up using components from several different candidates. Some examiners find pseudo-candidates harder to judge, others find them easier. Researchers tend to seek whole candidates who have a balanced profile of achievement upon each component of the assessment, because this makes the process of making judgements easier. Nonetheless such candidates are in practice rare (Sharaschkin & Baird, 2000), and it can be argued that limiting the sample to such candidates means that the sample is not representative of the population as a whole, where the average candidate will balance weakness in one area by strengths in another.

The issues become yet more complex when components contain question choice, and a candidate's opportunity to play to their strengths is maximised. One solution to this is to carry out comparability studies at component level, and not attempt the 'holistic' picture. Obviously this does not provide any indication of whether differences in components 'balanced out' to provide assessments which were comparable overall, but it does allow adjustments at component level to be made to bring components into line if this is the object of the exercise. Naturally it is only applicable where assessments have parallel components.

Comparing scripts

Cross-moderation meetings to compare candidates' work are usually held residentially over two or three days. Judges are asked to work individually, with scripts generally being rotated between examiners. Inevitably as they work examiners make comments about the nature of the work, or about scripts. Whilst such comments are discouraged, because the object is to retain individual judgements rather than group consensus, they are often useful in illuminating issues which later arise in the statistics. A plenary session, once the examiners have made their judgements, is a valuable method of eliciting an indication of expert evidence about the comparability of the assessments in question. At this point examiners who have carried out judgements about whole candidate's assessments can suggest whether there are irregularities between components which would not be evident in the overall statistics. Alternatively those who have judged individual components may comment upon the extent to which they believe the overall assessment is comparable.

Part II. A proper role for human judgement?

Introduction

Like so much of the whole subject area of standards and assessment, we are concerned in this paper to establish a number of definitions and distinctions:

- The difference between monitoring comparability and maintaining comparability.

- The use of human judgement in comparability versus and alongside statistical methods.
- The distinction between monitoring comparability and investigating comparability.

Maintaining comparability tends to take place during marking (and preparation for marking) and at the awarding meeting. During the marking process, human judgement drives the bulk of the process as co-ordination meetings and examiner's inherent knowledge of standards which have gone before enable them to continue to mark examination questions to a recognised standard for that syllabus. At the awarding meeting statistical influences are brought to bear – the performance of the current and previous cohort as a whole, syllabus pairs analyses, and to an extent forecast grades, although the latter brings us back to human judgement again.

It is possible to monitor comparability post–examination by conducting formal comparability studies. In this way it is possible to compare syllabuses which might not be compared routinely. Such studies are very often used simply to establish whether syllabuses are in line overall, but in order to act upon such findings it is necessary to provide more detail about where precisely within the syllabuses the differences lie. It is possible that syllabuses which appear to be in line when judged by overall comparability mask differences at component level which taken together even themselves out. This may not be an undesirable outcome, given that examinations by their nature are compensatory – a harder component compensated for by an easier component is acceptable as long as the overall standard is the same, but it is one that needs to be recognised. Where attempts are made to investigate with greater precision where exact differences are located (e.g. via component level comparability studies, or using Kelly's Repertory grid to establish differences in the specifications/curriculum) investigation of comparability is taking place.

Issues arising from the methodologies.

The ratification ("home and away") method was used with success for a number of years, but has now been largely superseded by the Thurstone pairs methodology. However it is still worth looking at the issues which affected this methodology as they remain pertinent to both current methodology and awarding.

It is not possible to make any estimate about the stability of the internalised standard, as utilised by the examiners. The methodology does much to ensure that examiners have an internalised standard (selecting examiners who work for that syllabus alone) and that this effect is maximised during the cross-moderation process (preparation work concentration upon the 'home' standard, and the judging of home scripts first to further establish the standard). Nonetheless there is little that can be done to measure the extent of the strength of the standard established, and whether it is the same for all examiners.

In order to convert judgements to statistics, assumptions are made that the categories chosen represent an equal interval scale, and that it is the same ordinal scale for each judge in question – clearly an issue beset with problems. Furthermore there is no evidence to be gleaned about the size of the middle (or borderline) category. A very wide interpretation of the acceptable borderline category by a number of judges would lead inevitably to good comparability – but maybe this was what the Boards intended?

Potentially the ratification method could have utilised forced choice – allowing examiners to only judge the scripts as better or worse than their borderline standard. We assume that this did not occur because it would have been at odds with the underlying assumptions driving the approach at the time.

The general adoption of the Thurstone method marked a change in the logic of such studies towards a method designed to uncover variation. Thurstone's methodology identifies consistent ordering and constructs a scale on which to measure it – it constructs an interval scale from the judgements. Thus comparing directly cancels out the internal standard.

The Thurstone method has practical advantages over the ratification method – it makes direct comparisons between the scripts under scrutiny, and home and away biases can be experimentally controlled for. It has a good reporting structure, and the pattern of both judge and script fit and misfit enables further investigation of the biases which may be taking place. However the practical application of this method moves away from the original concept of the methodology as described by Thurstone (1927). Judgements are not as instantaneous as Thurstone expected, and the extent to which this affects the validity of the results is as yet undetermined.

It is also as difficult with the Thurstone method as it is with the ratification method to translate the results to a meaningful scale. The size of each individual logit unit is not easy to determine, although the use of scripts known to be a fixed distance from the boundary can provide an approximate picture. Nonetheless even with these scripts we only know their place on a different scale (i.e. the original marking) and there may be many reasons why this does not correspond to the scale created by the Thurstone pairs exercise.

An investigation of bias

Critical to the ratification methodology was the notion of systematic bias on the part of all the examiners to their home Board. This remains a concern of the Thurstone methodology, and is a reason why balanced teams of examiners are used wherever possible. Circumstances, however, have changed and the existence and significance of 'home bias' cannot be assumed, any more than the absence of any other sort of bias can be. An important advantage of the Thurstone approach over the ratification methodology is that it is, in statistical terms, a *strong* model, defined at the individual level and modeling each judgement explicitly, rather than one that is only defined at the group level treating all the judges from one board as equivalent. Such a strong model makes it possible to investigate questions of bias from any source whatever.

Thurstone described how to analyse data collected in this way in the 1920s. He assumed that a particular judge would assign a value to a particular object in a way that could be represented by a random variable, normally distributed around the "true" value of that object for that judge. If the logistic distribution is substituted for the normal distribution Thurstone's model becomes identical to that described by Georg Rasch (1960, 1966). Standard Rasch model programs like FACETS (Linacre, 198?) can be used to generate estimates of the standard of each script and, importantly, measures of the consistency of,

or amount of surprise in, each judgement that is made. The analyses reported here, however, used locally written software.

To illustrate, Appendix A contains the analysis of one recent comparability study, which involved three examinations. Three judges from each examination took part, and eight scripts were chosen from close to the grade boundary in each, though it turned out that one script was flawed, and only twenty three were actually used.

There are 253 (23 x 22 / 2) possible comparisons involving 23 scripts, and in this study the judges each made about one-third of these comparisons. Note that this method is not affected by the absence of two-thirds of the possible data; compared to the ratification method it is very robust. In practice researchers ensure that each judge makes a fairly representative sample of the possible comparisons, even though in principle the method is not affected by the representativeness of each sample.

The analysis first reports all of the comparisons in a two-way table, and then a summary of the 'success rate' of each script, in a sort of league table – indeed this Thurstone / Rasch analysis is an excellent way of analysing data from the likes of football matches, and as good as any way of forecasting the results of forthcoming matches.. This is the starting point for the analysis.

Standard 'maximum likelihood' methods are then used to estimate the relative worth (called 'parameter estimate') of each script, using the explicit model that the odds of script S 'beating' script T in a comparison are the exponential of ValueS – ValueT.

The estimated values are reported, first in a formal table, then in the form of a chart referred to earlier. In this case the scripts from the three examinations have been distinguished as **bold**, *italics* and <u>underline</u>. It looks as if examination **X** may be at a higher standard than the other two, a suggestion that is quickly checked by a one-way analysis of variance. Here the F value for the test is calculated to be 3.98 (prob = 0.034) indicating – just – a significant variation in the standards of the three exams. The usual procedure is to use a regression analysis to indicate how many marks adrift the standards are.

The real power of the Thurstone approach, though, follows this analysis. The program goes on to analyse how well the data fit the Thurstone model. The primary assumption of the analysis, just as it is the primary assumption of awarding, is that judges differ *only* in how high a standard they set. In Thurstone's method each judge's standard cancels out when a comparison is made, and any misfit between data and model must indicate a difference of some other kind. This is studied by analysing the *residuals* that remain after the model is fitted to the data. Whenever one script is judged to have a higher value than a competing one it is expected to 'win' the comparison: if it does the residual will be small, but if it loses then the residual for that judge's comparison of those two scripts will be large. Any patterns of large residuals will indicate some sort of bias – in this context *bias* simply means a source of systematic, rather than random, variability.

The basic analysis summarise and evaluates these residuals for each script and each judge, identifying 'misfitting' cases. When a script misfits, it is normal to remove it, since it is easy to conceptualise a misfitting script as one that is unrepresentative of the exam as a whole. It is less obvious what to do about a misfitting judge, since varying

interpretations of the trait that is being assessed seem a more authentic part of the assessment process. In fact, however, removing one or two judges from the analysis makes little difference to the estimates of the relative value of scripts, and so does not have much impact on the main analysis.

Appendix B is a custom-designed analysis of the same residuals to investigate the hypothesis that judges exhibit bias in favour of home scripts. In each table the key column is the one headed **z stat**; any value of z stat greater than about 2 indicates a statistically significant amount of judgement bias. Table A repeats the basic analysis, and indicates that one judge – Judge 5 – is significantly biased in some way. The general lack of evidence for home (or away) bias in this table should be remembered in what follows.

Table B repeats this analysis separately for judgements involving one home script, **Home**, and those involving two away scripts, **Away**. The z statistics confirm that Judge 5 is more biased in home comparisons than in those only involving away scripts. Some other judges, especially Judges 3 and 7, show a slight tendency towards home bias, while others, notably Judges 1 and 6, are significantly more unbiased in home comparisons than other judges. The bottom line, for all home and away comparisons, shows again that there is no evidence for a general home/away bias.

Finally, Table C splits the analysis further, separating home wins from home losses. At this level of detail some statistics are based on very small numbers. The generally higher level of exam X scripts means that Judges 1, 6 and 8, the X examiners, are not expected to make many judgements that their home scripts should lose. They appear very *unbiased* as a result. Because these statistics must even out over the whole data set most of the others show some indication of home bias.

In summary, there is no convincing evidence here of a significant systematic home bias, though one judge may be affected by it, but when the variations are probed further what misfit there is can mostly be put down to a non-significant home bias effect.

Discussion

A similar analysis could be carried out for any other hypothetical source of systematic influence on the comparative judgement. It might for instance be thought that the sex of the student, or of the judge, or the quality of the handwriting could influence judgements, and these hypotheses could be investigated in the same way.

Thurstone developed this method as a way of constructing *de novo* a scale for measuring the quality of objects whose value could be judged instantaneously; in our context this is clearly not the case, as judges may take several minutes to gauge the worth of one script (which may be a set of work, as mentioned earlier). In addition, and as a consequence of this, it seems possible that the judgements we analyse may not be as independent as he assumed. The general success of the analysis, however, suggests that these concerns do not invalidate the method, and we should continue to consider how best to apply it to validate our assessment processes.

The analysis shows the power of the Thurstone approach as a means of post-hoc investigation of factors that might cause unfairness in awarding. It could be applied within a single examination, as well as between two or more exams as here. The analysis is very simple and quick, raising the possibility that it could be used during the marking

process, or after marking but before awarding - so long as a certain amount of double marking of scripts is possible.

But it is most clearly applicable after the awarding is complete, as a check that standards are being maintained. In our presentation at the seminar we will consider further the proper role for judgement in the assessment process – and how judgement could be integrated with statistical methods of maintaining standards.

References

Bell J. F., Bramley T. and Raikes N. (1997) *Standards in A level Mathematics 1986-1996.* A paper presented at the British Educational Research Association Annual Conference. September 11-14 1997. York.

Elliott, G and Greatorex, J. (2002) A fair comparison? The evolution of methods of comparability in national assessment. *Educational Studies.* 28(3) 2002. 253-264.

Forster, M. and Gray, E. (2000) *Impact of independent judges in comparability studies conducted by awarding bodies*. A paper presented at the British Educational Research Association Annual Conference, Cardiff University, September 7-10.

Houston J. G. (1975*) Report of the inter-board cross-moderation study in English Literature at Ordinary Level.* AEB. 1975.

Linacre, JM. (1989) Many-Facet Rasch Measurement. MESA Press, University of Chicago.

Massey A. and Newbould C.(1977) *Comparability by cross moderation: A methodological retreat or a conceptual advance?* A paper prepared for the annual conference of the British Educational Research Association. Nottingham. 6-9 September 1977.

Rasch, G. (1960, 1966) Probabilistic Models for some Intelligence and Attainment Tests. National Institute for Educational Research, Copenhagen, 1960/ University of Chicago Press, 1966.

Sharaschkin A. and Baird J. (2000) The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*. 26(3) June 2000, 33-357.

Thurstone L. L. (1927) The law of comparative judgement. *Psychological Review.* 34 273-286.

Appendix A : Parameter analysis for the sample study

RASCH ANALYSIS using the PAIRED COMPARISON model *** Sample study
Table of Comparisons:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	.	.	.	.	.	.	.	.	1	1	2	.	2	1	.	4	4	1	3	2	2	3	3
2	.	.	.	.	.	.	.	.	.	.	.	1	2	2	1	2	1	.	.	2	3	1	2
3	.	.	.	.	.	.	.	.	1	3	2	1	.	1	1	1	2	.	1	2	4	1	2
4	.	.	.	.	.	.	.	.	2	1	.	1	.	1	.	2	1	.	2	3	4	2	4
5	.	.	.	.	.	.	.	.	2	1	1	1	.	1	1	1	.	.	3	3	4	4	2
6	.	.	.	.	.	.	.	.	.	2	.	.	1	2	.	2	4	1	2	2	.	.	1
7	.	.	.	.	.	.	.	.	3	3	2	.	2	1	3	6	1	3	2	2	7	1	5
8	.	.	.	.	.	.	.	.	.	2	1	1	.	2	2	3	1	.	2	3	3	.	3
9	4	2	2	2	3	4	2	3	.	.	.	.	.	.	.	5	4	1	3	2	3	2	4
10	2	6	5	3	1	4	1	3	.	.	.	.	.	.	.	5	1	.	1	2	2	.	2
11	2	6	3	2	4	2	1	6	.	.	.	.	.	.	.	4	3	.	3	3	4	1	2
12	7	9	5	3	5	5	5	4	.	.	.	.	.	.	.	6	7	2	5	5	3	2	6
13	2	4	4	4	3	1	.	4	.	.	.	.	.	.	.	4	4	2	2	3	1	2	3
14	4	4	4	2	5	1	1	3	.	.	.	.	.	.	.	2	2	1	3	3	4	1	4
15	6	4	2	5	2	4	1	3	.	.	.	.	.	.	.	3	2	2	3	3	2	1	4
16	.	2	1	1	1	1	.	.	1	2	.	.	.	.	1	.	.	.	.	.	.	.	.
17	3	1	3	2	3	1	.	1	2	4	2	.	.	.	1	.	.	.	.	.	.	.	.
18	2	3	1	6	7	4	2	6	6	4	.	1	1	2	3	.	.	.	.	.	.	.	.
19	2	.	5	2	.	2	2	1	3	.	1	.	2	1	.	.	.	.	.	.	.	.	.
20	1	2	3	2	1	1	1	2	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.
21	4	1	.	.	.	2	.	.	3	2	2	1	1	1	1	.	.	.	.	.	.	.	.
22	2	8	2	5	4	5	2	5	2	1	4	4	1	3	3	.	.	.	.	.	.	.	.
23	2	1	1	3	2	2	.	1	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.

Row object "beats" Column object (has more of the trait):
e.g. Object 1 beats Object 9 1 time; Object 9 beats Object 1 4 times

RASCH ANALYSIS using the PAIRED COMPARISON model *** Sample study
Wins and Losses for each object

Object		Wins	Losses	Comparisons	%
1		Y1	29	43	72	40.3
2		Y2	17	53	70	24.3
3		Y3	22	41	63	34.9
4		Y4	23	42	65	35.4
5		Y5	24	41	65	36.9
6		Y6	17	39	56	30.4
7		Y7	41	18	59	69.5
8		Y8	23	42	65	35.4
9		X1	46	26	72	63.9
10		X2	38	29	67	56.7
11		X3	46	18	64	71.9
12		X4	79	11	90	87.8
13		X5	43	12	55	78.2
14		X7	44	18	62	71.0
15		X8	47	17	64	73.4
16		Z1	10	50	60	16.7
17		Z2	23	37	60	38.3
18		Z3	48	13	61	78.7
19		Z4	21	35	56	37.5
20		Z5	15	40	55	27.3
21		Z6	18	46	64	28.1
22		Z7	51	21	72	70.8
23		Z8	14	47	61	23.0

Total number of comparisons = 739

RASCH ANALYSIS using the PAIRED COMPARISON model *** Sample study
Summary of Pair Iteration Process
Parameter Estimate Shifts

Iter	1	RMS: 0.7996	Max: 1.3272	Object:	X4	Var: 0.6303
Iter	2	RMS: 0.1397	Max: 0.3456	Object:	X4	Var: 0.7740
Iter	3	RMS: 0.0227	Max: 0.0651	Object:	X4	Var: 0.8127
Iter	4	RMS: 0.0026	Max: 0.0097	Object:	Y7	Var: 0.8164
Iter	5	RMS: 0.0002	Max: 0.0005	Object:	Y3	Var: 0.8168

Appendix A : Parameter analysis for the sample study

RASCH ANALYSIS using the PAIRED COMPARISON model *** Sample study
Estimates of object parameters

Number	Parameter	Object Name
1	-0.273	Y1
2	-0.913	Y2
3	-0.600	Y3
4	-0.546	Y4
5	-0.213	Y5
6	-0.732	Y6
7	1.111	Y7
8	-0.394	Y8
9	0.411	X1
10	-0.213	X2
11	0.579	X3
12	1.812	X4
13	1.029	X5
14	0.610	X7
15	0.877	X8
16	-1.511	Z1
17	-0.357	Z2
18	1.576	Z3
19	-0.320	Z4
20	-1.017	Z5
21	-0.834	Z6
22	1.073	Z7
23	-1.157	Z8

Appendix A : Parameter analysis for the sample study

```
RASCH ANALYSIS using the PAIRED COMPARISON model    *** Sample study
Plot of Parameter Estimates

    | 2|
    |  |       X4
    |  |
    |  |       Z3
    |  |
    |  |       Y7
    | 1|       X5     Z7
    |  |       X8
    |  |       X7
    |  |       X3
    |  |       X1
    |  |
    | 0|
    |  |       Y5     X2
    |  |       Y1     Y8     Z2     Z4
    |  |       Y4
    |  |       Y3     Y6
    |  |       Y2     Z6
    |-1|       Z5
    |  |       Z8
    |  |
    |  |       Z1
    |  |
    |  |
    |-2|
```

```
RASCH ANALYSIS using the PAIRED COMPARISON model    *** Sample study
Misfitting judgements (Standardised residual > 2.0)

Judge X J1 says:  2: "Y2"  beats 15: "X7"
Calculated probability is  0.821    Standardised residual is  2.14
Judge X J1 says:  3: "Y3"  beats 23: "Z7"
Calculated probability is  0.158    Standardised residual is  2.31
Judge X J1 says:  4: "Y4"  beats 23: "Z7"
Calculated probability is  0.165    Standardised residual is  2.25
Judge Y J2 says:  2: "Y2"  beats 13: "X5"
Calculated probability is  0.875    Standardised residual is  2.64
Judge Y J2 says:  8: "Y8"  beats 12: "X4"
Calculated probability is  0.901    Standardised residual is  3.01
Judge Y J3 says:  2: "Y2"  beats 13: "X5"
Calculated probability is  0.875    Standardised residual is  2.64
Judge Y J3 says:  4: "Y4"  beats 12: "X4"
Calculated probability is  0.914    Standardised residual is  3.25
Judge Z J4 says:  3: "Y3"  beats 12: "X4"
Calculated probability is  0.918    Standardised residual is  3.34
Judge Z J4 says:  6: "Y6"  beats 13: "X5"
Calculated probability is  0.853    Standardised residual is  2.41
Judge Z J4 says:  4: "Y4"  beats 23: "Z7"
Calculated probability is  0.165    Standardised residual is  2.25
Judge Z J5 says:  2: "Y2"  beats 12: "X4"
Calculated probability is  0.939    Standardised residual is  3.91
Judge Z J5 says:  3: "Y3"  beats 16: "X8"
Calculated probability is  0.814    Standardised residual is  2.09
Judge Z J5 says:  5: "Y5"  beats 12: "X4"
Calculated probability is  0.883    Standardised residual is  2.75
Judge Z J5 says: 17: "Z1"  beats  9: "X1"
Calculated probability is  0.872    Standardised residual is  2.61
Judge Z J5 says: 22: "Z6"  beats 11: "X3"
Calculated probability is  0.804    Standardised residual is  2.03
Judge Z J5 says: 22: "Z6"  beats 12: "X4"
Calculated probability is  0.934    Standardised residual is  3.76
Judge Z J5 says: 22: "Z6"  beats 13: "X5"
Calculated probability is  0.866    Standardised residual is  2.54
Judge Z J5 says: 24: "Z8"  beats 11: "X3"
Calculated probability is  0.850    Standardised residual is  2.38
Judge Z J5 says: 20: "Z4"  beats  7: "Y7"
Calculated probability is  0.807    Standardised residual is  2.05
```

Appendix A : Parameter analysis for the sample study

```
Judge Z J7 says: 17: "Z1"  beats 16: "X8"
Calculated probability is  0.916    Standardised residual is  3.30
Judge Z J7 says: 22: "Z6"  beats 11: "X3"
Calculated probability is  0.804    Standardised residual is  2.03
Judge Z J7 says: 21: "Z5"  beats  7: "Y7"
Calculated probability is  0.894    Standardised residual is  2.90
Judge X J8 says:  1: "Y1"  beats 19: "Z3"
Calculated probability is  0.136    Standardised residual is  2.52
Judge X J8 says:  6: "Y6"  beats 19: "Z3"
Calculated probability is  0.090    Standardised residual is  3.17
Judge Y J9 says:  2: "Y2"  beats 15: "X7"
Calculated probability is  0.821    Standardised residual is  2.14
Judge Y J9 says:  2: "Y2"  beats 16: "X8"
Calculated probability is  0.857    Standardised residual is  2.45
Judge Y J9 says: 22: "Z6"  beats 15: "X7"
Calculated probability is  0.809    Standardised residual is  2.06
Judge Y J9 says: 22: "Z6"  beats 16: "X8"
Calculated probability is  0.847    Standardised residual is  2.35
Judge Y J9 says: 20: "Z4"  beats  7: "Y7"
Calculated probability is  0.807    Standardised residual is  2.05
Judge Y J9 says:  2: "Y2"  beats 23: "Z7"
Calculated probability is  0.121    Standardised residual is  2.70
-----------------------------------------------------------------

  30 comparisons ( 4.1% ) exceed the criterion of  2.00
```

Appendix A : Parameter analysis for the sample study

RASCH ANALYSIS using the PAIRED COMPARISON model *** Sample study
Summary of Fit Statistics for Judges

Number	Name	Parameter	MeanSq	Unwtd-t	Wtd-t
1	X J1	-----	0.772	-1.724	-1.955
2	Y J2	-----	1.032	-0.004	0.287
3	Y J3	-----	1.110	0.465	0.870
4	Z J4	-----	1.063	0.107	0.505
5	Z J5	-----	1.261	2.794	2.228
6	X J6	-----	0.666	-2.734	-2.944
7	Z J7	-----	0.963	-0.128	-0.266
8	X J8	-----	0.958	-0.462	-0.302
9	Y J9	-----	1.156	0.628	1.323
		Mean:	0.998	-0.117	-0.028
		S.D.:	0.175	1.452	1.508

RASCH ANALYSIS using the PAIRED COMPARISON model *** Sample study
Summary of Fit Statistics for Objects

Number	Name	Parameter	MeanSq	Unwtd-t	Wtd-t
1	Y1	-0.273	1.071	0.453	0.664
2	Y2	-0.913	1.043	0.559	0.313
3	Y3	-0.600	1.070	0.773	0.641
4	Y4	-0.546	0.978	0.044	-0.129
5	Y5	-0.213	1.027	0.250	0.255
6	Y6	-0.732	1.034	0.125	0.285
7	Y7	1.111	0.959	-0.453	-0.239
8	Y8	-0.394	0.849	-0.891	-1.186
9	X1	0.411	1.002	-0.008	0.054
10	X2	-0.213	0.970	-0.376	-0.323
11	X3	0.579	0.923	-0.557	-0.505
12	X4	1.812	1.011	0.058	0.126
13	X5	1.029	1.068	0.270	0.393
14	X7	0.610	1.007	0.029	0.099
15	X8	0.877	0.939	-0.024	-0.324
16	Z1	-1.511	0.949	-0.452	-0.166
17	Z2	-0.357	0.973	-0.690	-0.257
18	Z3	1.576	0.860	-0.870	-0.749
19	Z4	-0.320	1.060	0.386	0.600
20	Z5	-1.017	0.931	-0.621	-0.481
21	Z6	-0.834	1.169	1.493	1.304
22	Z7	1.073	1.204	1.174	1.480
23	Z8	-1.157	0.912	-1.019	-0.528
		Mean:	0.959	-0.014	0.055
		S.D.:	0.216	0.624	0.597

Appendix B : Analysis for Home/Away bias in the sample study

Table A: Overall Fit Statistics

Judge		ΣWz2	ΣW	WMS	WMS1/3	Σ(K-W2)	q	z stat	Mean Res	N
X	**1**	10.84	14.04	0.77	0.92	3.03	0.12	**-1.96**	0.31	81
Y	**2**	14.16	13.72	1.03	1.01	3.21	0.13	**0.29**	0.34	81
Y	**3**	15.54	14.00	1.11	1.04	3.21	0.13	**0.87**	0.36	81
Z	**4**	14.37	13.52	1.06	1.02	3.33	0.13	**0.50**	0.34	81
Z	**5**	20.75	16.46	1.26	1.08	3.28	0.11	**2.23**	0.41	90
X	**6**	9.18	13.78	0.67	0.87	3.08	0.13	**-2.94**	0.29	79
Z	**7**	14.15	14.70	0.96	0.99	3.23	0.12	**-0.27**	0.34	84
X	**8**	13.30	13.89	0.96	0.99	2.96	0.12	**-0.30**	0.34	80
Y	**9**	16.96	14.67	1.16	1.05	2.88	0.12	**1.32**	0.39	82
	All	129.27	128.78	1.00	1.00	28.21	0.04	**0.11**	0.349	739

Table B: Home & Away

		Home									Away								
Judge		ΣWz2	ΣW	WMS	W1/3	Σ(K-W2)	q	z stat	M Res	N	ΣWzSq	ΣW	WMS	W1/3	Σ(K-W2)	q	z stat	M Res	N
X	**1**	5.12	7.81	0.66	0.87	2.0	0.18	**-2.09**	0.27	48	5.72	6.24	0.92	0.97	1.0	0.16	**-0.48**	0.36	33
Y	**2**	10.49	9.70	1.08	1.03	1.9	0.14	**0.61**	0.38	53	3.68	4.02	0.91	0.97	1.3	0.28	**-0.22**	0.28	28
Y	**3**	11.98	9.87	1.21	1.07	1.8	0.14	**1.49**	0.41	53	3.56	4.14	0.86	0.95	1.3	0.28	**-0.43**	0.27	28
Z	**4**	9.32	8.95	1.04	1.01	2.0	0.16	**0.31**	0.34	53	5.05	4.57	1.11	1.03	1.2	0.25	**0.50**	0.34	28
Z	**5**	15.44	11.90	1.30	1.09	2.1	0.12	**2.24**	0.43	64	5.31	4.55	1.17	1.05	1.1	0.23	**0.76**	0.38	26
X	**6**	6.89	9.55	0.72	0.90	2.1	0.15	**-1.97**	0.30	54	2.28	4.23	0.54	0.81	0.9	0.23	**-2.34**	0.26	25
Z	**7**	10.61	8.69	1.22	1.07	1.9	0.16	**1.33**	0.38	51	3.54	6.00	0.59	0.84	1.2	0.19	**-2.53**	0.29	33
X	**8**	6.96	8.47	0.82	0.94	1.9	0.17	**-1.09**	0.31	50	6.34	5.41	1.17	1.05	0.9	0.18	**0.94**	0.39	30
Y	**9**	12.19	10.69	1.14	1.04	1.8	0.13	**1.10**	0.40	57	4.77	3.98	1.20	1.06	1.0	0.25	**0.81**	0.35	25
	All	89.01	85.63	1.04	1.01	17.96	0.05	**0.80**	0.362	483	40.26	43.15	0.93	0.98	10.25	0.07	**-0.90**	0.326	256

Table C: Home Wins and Losses

		Home Win	Home Loss

Appendix B : Analysis for Home/Away bias in the sample study

	Judge	ΣWz2	ΣW	WMS	W1/3	Σ(K-W2	q	z stat	M Res	N	ΣWzSq	ΣW	WMS	W1/3	Σ(K-W2	q	z stat	M Res	N
X	**1**	3.25	6.59	0.49	0.79	1.84	0.21	**-3.00**	0.23	42	1.88	1.22	1.54	1.15	0.20	0.36	**1.39**	0.52	6
Y	**2**	8.71	6.13	1.42	1.12	1.04	0.17	**2.29**	0.46	32	1.77	3.57	0.50	0.79	0.86	0.26	**-2.32**	0.25	21
Y	**3**	9.25	5.81	1.59	1.17	0.86	0.16	**3.20**	0.52	29	2.73	4.05	0.67	0.88	1.01	0.25	**-1.42**	0.00	24
Z	**4**	7.22	5.87	1.23	1.07	0.91	0.16	**1.37**	0.44	30	2.10	3.07	0.68	0.88	1.15	0.35	**-0.90**	0.23	23
Z	**5**	12.05	7.60	1.59	1.17	1.26	0.15	**3.42**	0.49	40	3.39	4.31	0.79	0.92	0.91	0.22	**-0.97**	0.32	24
X	**6**	3.29	6.99	0.47	0.78	1.77	0.19	**-3.43**	0.24	42	3.60	2.56	1.41	1.12	0.35	0.23	**1.64**	0.51	12
Z	**7**	9.11	5.80	1.57	1.16	1.00	0.17	**2.88**	0.50	30	1.50	2.89	0.52	0.80	0.97	0.34	**-1.62**	0.21	21
X	**8**	4.76	6.98	0.68	0.88	1.77	0.19	**-1.82**	0.27	43	2.20	1.49	1.48	1.14	0.21	0.31	**1.47**	0.53	7
Y	**9**	7.78	5.03	1.55	1.16	0.61	0.16	**3.08**	0.53	24	4.41	5.66	0.78	0.92	1.24	0.20	**-1.16**	0.31	33
	All	65.43	56.80	1.15	1.05	11.07	0.06	**2.49**	0.392	312	23.58	28.83	0.82	0.94	6.89	0.09	**-2.10**	0.306	171

Summary of data:

	Prob	Resid	Var	z	z2	Wz2	Kurtosis	K-W2
Mean	0.65	0.349	0.17	0.82	**0.99**	0.17	0.07	0.04
SD	0.23	0.23	0.06	0.56	1.61	0.20	0.01	0.02