Constructing scales of complex objects using comparative judgement methods

Ian Jones and Ben Davies

Loughborough University, UK

UCL, UK

Most recent edits: Ben Davies [04/03/2021]

Abstract

* DESIGN CONSIDERATIONS
* EVALUATING THE SCALE
  + Approaches to reliability (SSR, inter-rater)
  + Approaches to validity
* DISCUSSION
  + A simple idea, with many potential applications across the social sciences
  + Common objections: it’s black box; it’s inherently norm-referenced; it’s inefficient

Keywords: Comparative Judgement, Measurement, Reliability, Validity

**Introduction**

A common barrier to advancing our knowledge of the social sciences relates to measurement. Researchers commonly need to position complex objects, which might be very different to one another, on a linear scale. The scale represents a construct, which is typically an attribute or theoretical idea that is important but nebulous and therefore difficult to define or operationalise. Examples of such constructs are social attitudes, conceptual understanding of mathematics, beauty, problem solving skills and so on.

There are numerous methods for constructing linear scales of nebulous constructs in the social sciences. For example, in fields such as psychology, health and education typical methods include clinical interviews (Ginsburg, 1997) and developing psychometrically robust objective tests (e.g. Burckhardt & Anderson, 2003), but these methods tend to be resource and time-intensive, especially for the types of nebulous constructs of interest here. In other fields, such as marketing and environmental studies, choice modelling methods are widespread (Johnston et al., 2017; Pearce, Atkinson & Mourato, 2016), in which participants directly select or rank complex objects. Choice modelling methods were developed in part from Thurstone’s (1927a) work, and Thurstone’s method based on paired comparisons, referred to here as comparative judgement, remains in use today. In recent decades many education scholars have embraced comparative judgement techniques for improving the reliability and validity of educational assessments (e.g. Bramley, 2007; Jones, Wheadon, Humphries & Inglis, 2016; Pollitt, 2012). Outputs from this work include online platforms for running comparative judgement experiments (see SECTION), robust methods for investigating the reliability and validity of comparative judgement outcomes (e.g. Jones, Bisson, Gilmore & Inglis, 2019), and guidance for designing comparative judgement experiments (e.g. Verhavert, Bouwer, Donche & De Maeyer, 2019). The purpose of this paper is to provide an overview of comparative judgement methods, and to synthesise and generalise research findings in order to offer study design guidance to researchers across the social sciences who want to construct scales of complex objects.

First, we offer two example applications of comparative judgement methods, one drawn from social psychology and the other from education. We then provide a full description of the underpinning theoretical framework derived from Thurstone’s (1927a) Law of Comparative Judgement, and provide guidelines and considerations for researchers conducting comparative judgement experiments.

**Applications**

To acquaint the reader with the kinds of research questions that comparative judgement experiments can address, and the kinds of answers it can provide, we will consider two examples from the literature. One is a classic experiment conducted by Thurstone in the early 20th century, the other is a contemporary application from our own work on evaluating qualification standards.

**Measuring social values**

Thurstone (1927b) demonstrated his novel comparative judgement method by applying it to the measurement of social values, and in particular the judged seriousness of offences such as arson, homicide and rape. He was interested in “qualitative judgments of a rather intangible sort, loaded usually with personal opinion, bias, and even strong feeling, and regarded generally as the direct antithesis of quantitative measurement” (p. 398) and asked if it is “possible to reduce these qualitative judgments about the relative seriousness of difference offences to a quantitative basis?” (p. 385). There were 19 offenses of interest, and every possible pairing (*n*(*n* - 1)/2 = 171) was printed as a list and distributed to 266 undergraduate participants. Each participant was instructed to underline the offence in each pairing that they thought was the “more serious”. The pairwise decisions were statistically modelled using the mathematics set out in Thurstone (1927a) which resulted in a unique “scale value”, and standard deviation of the scale value, for each offense (we describe process for statistically modelling comparative judgement decision data in Section XX). These values were then used to construct a scale from least serious to most serious. For interest, the offences judged most serious were rape and homicide, which had very similar scale values (and standard deviations) of 3.275 (0.630) and 3.156 (0.682) respectively; the third most serious offence was seduction, which was much further down the scale with a relative scale value of 2.273 (0.438).

Thurstone was clear about that scale values represented “seriousness as judged” and that the values “may be quite wrong when looked at from the standpoint of objective checks or standards” (p. 384). (We discuss methods for externally validating scales constructed from comparative judgement experiments later in the article.) Thurstone did not comment that his scale is necessarily a cultural product particular to the population of undergraduates who participated in the study, although he did provide a sheet of definitions to the participants after discovering in a pilot study that they did not all know what all of the terms meant. The cultural particularness of the scale became evident when other researchers replicated his study decades later and produced scales in which the offences were in relatively different positions compared to Thurstone’s scale (e.g. Stone, 2000). It is therefore important that contemporary researchers are clear about what can and what cannot be inferred when interpreting scale values from their own comparative judgement experiments.

**Evaluating qualification standards**

Comparative judgement methods are common in studies that seek to evaluate standards in qualifications over time and across jurisdictions or institutions. For example, Jones, Wheadon, Humphries and Inglis (2016) studied changes over five decades in A-level Mathematics, a qualification in England and Wales that is widely considered a gatekeeper for entry to many university undergraduate programmes. Jones et al. acquired a historic archive of examination papers and 66 graded candidate scripts ranging from 1964 to 2012, which they split into 546 individual question responses. They retyped the examination questions and re-scribed the question responses in order to standardise the font and handwriting, and to remove examiners’ marks and comments. Twenty experts (mathematics PhD students) made a total of 5,000 comparative judgements on the individual question responses (see Figure X) and were asked to decide, for each pairing, which candidate was “the better mathematician”. Unlike Thurstone’s (1927b) study, only a subset of all possible pairings were judged: 5,000 out of 148,785 possible pairings, a mere 3.4%. The ability to produce reliable scales using only a fraction of all possible pairings is a key development over the past two decades that has rendered comparative judgement a viable method for large numbers of complex objects (Pollitt, 2012). We discuss the number of judged pairings required to construct valid and reliable scales later.

Graphical user interface, text, application

Description automatically generated

Figure X: Example from BERJ 2016 study (original image, no acknowledgement required)

Jones at al. statistically modelled the pairwise decisions (see Section X) to produce a scale value for each question response, and these values were averaged across each grade for each examination paper to produce the graph reproduced in Figure X. Regression analysis led to the conclusion that a grade B in 1996 or 2012 was perceived by the expert judges as reflecting the same achievement as a grade E in 1964 or 1968, suggesting a decline in standards since the 1960s that had stabilised since the mid 1990s. Unlike Thurstone (1927b), Jones et al. (2016) were able to provide some external validation to their findings because they had access to the original grades that had been assigned to the scripts by examiners. As seen in Figure X, at each given year scripts graded A received a higher mean scale value than scripts at grade B which received a higher mean scale value than scripts at grade E (only these three grades were available at each year, and grade B was missing for 1996). Conversely, and like Thurstone, the scale is a cultural product particular to the population of experts who conducted the judging (contemporary mathematics PhD students from a variety of countries). Should researchers repeat the study in future decades with a different population of expert judges, the scale may not replicate. We return to the issue of different populations judging the same complex objects later MAKE SURE WE DO!



Figure X: Reproduced from BERJ 2016 (p. 552)

**THE LAW OF COMPARATIVE JUDGEMENT**

To equip the social scientist with a comprehensive guide to comparative judgement, it is necessary to provide a complete overview of the statistical machinery underpinning the research method. In this section, we unpack the technical details of comparative judgment with historical and theoretical commentary along the way.

At its core, comparative judgment is a tool for estimating subjective human perceptions by quantitative values indicative of those perceptions. We first discuss the theoretical model of human perception underpinning the translation of subjective perceptions to pairwise comparisons. We then consider the computations necessary to transform the pairwise comparisons (referred to as judgments) to numerical estimates of the perceived quality of each response.

#### Using mathematics to model human perceptions

In this paper, we are interested in estimating the perceived volume of a given social construct present in a series of artefacts or responses. Before discussing the role of pairwise comparisons in this process, we need a mathematical model of the way judges perceive these artefacts in isolation. This model of perception can then be used in conjunction with empirical judgment data to generate an estimate of the likelihood of a judge choosing one artefact as having more of a given construct than another over another, eventually, a numerical estimate of perceived volume for each artefact.

Each time a judge encounters an artefact, *A*, it is perceived as lying somewhere on a continuum of merit. I say is the merit assigned to artefact *A* in encounter *i*. Thurstone (1927) called the process of assigning that merit the *discriminal process.* This may be unstable in time as a judge may perceive the volume of the construct in question as different in different encounters (perhaps influenced by mood, time of day, or other artefacts that judge has recently encountered)[[1]](#footnote-1).

The are assumed to be normally distributed, with standard deviation, , about the mean, . We use here to connote the collective `value' assigned to the artefact *A*, via the various encounters with *A*. We call these normal distributions discriminal dispersions (Bramley 2007). See Figure X, showing the discriminal dispersions for artefacts *A* and *B* with distributions and

![Diagram

Description automatically generated with medium confidence]()

Figure X. Illustration of two overlapping distributions for artefacts A and B. Labels on the X-axis correspond to encounters of a given artefact. Adapted from Bramley (2007).

When a judge compares two artefacts, Thurstone imagined that it is the values resulting from the discriminal processes that are compared. When artefacts *A* and *B* are compared, if a judge perceives *A* as having ‘more’ of the given construct that *B*, then the judge will assert *A* ‘beats’ *B*. Algebraic, we represent this as . Notice that in Figure X, while the discriminal dispersion for *A* is further along the merit continuum than *B*, the overlapping distributions mean that it is possible, based on some encounters of the two artefacts, for a judge to assert `*B* beats *A*'.

To evaluate the likelihood of the two outcomes, we consider the distribution given by the difference of the discriminal dispersions, with standard deviations and Call this new distribution the paired discriminal dispersion, with standard deviation

where is the correlation between discriminal dispersions[[2]](#footnote-2). The new distribution is also normal, with mean , the difference (see Figure Y). ![Diagram

Description automatically generated]()

Figure Y. An illustration of the distribution centered at used to estimate the likelihood of A beats B (the shaded area to the right of zero) and B beats A (to the left of zero). Adapted from Bramley (2007).

As is shown in Figure Y, for a given comparison, the probability that *A* beats *B* is the proportion of the distribution where . This is the area to the right of zero under the curve with mean , and is determined by the *z-*score of zero in the distribution . This is used to state a general form of Thurstone's law of comparative judgment as , where is precisely the *z-*score of zero.

The most general version of the law of comparative judgment applies only to comparisons of a single pair of artefacts. To generate a model more useful for practical application, Thurstone proposed a series of five cases, each imposing stricter assumptions than the last. The fifth and final case assumes that every object has the same discriminal dispersion, call it , and that all dispersions are uncorrelated, i.e. . This results in = and allows the simplification . The denominator here is constant and can be considered an arbitrary unit of measurement so, without losing information, we can equivalently state the law of comparative judgment as

Our unit of measurement imbues a particular meaning on the scores produced in the eventual model, allowing scores to be interpreted as standard deviations from the mean.

From Thurstone's assumptions, we can approximate the likelihood that ‘*A* beats *B*' by considering the area to the right of zero, under the standard normal curve centered at

This is given

#### From a model to a measure

Recall that our goal is to estimate the relative volume of constructs in given artefacts using binary pairwise comparisons. So far, we have a way of mathematising a single comparison, *A* vs *B*, and a probability model for estimating the likelihood of *A* beating *B* and vice versa. Before we can begin the process of assigning scores to artefacts, we have two problems to solve. The first is that the integral of the normal distribution function famously has no analytical solution. The second is that the above function is dependent on and .

The first problem was solved by Andrich (1978), who proposed replacing the normal distribution with a logistic one with near-identical outputs:  
.

By setting an arbitrary scaling parameter, to , Andrich's new model generates near-identical outputs for the two distributions (see Figure Z). However, the values generated by Thurstone's model have no particular importance, so for simplicity it is standard to set ,

resulting in the simpler logistic model,  
.

![Diagram

Description automatically generated]()

Figure Z. A visual comparison of the logistic (Andrich) and Normal (Thurstone) models. Adapted from Bramley (2007).

The value of Andrich's new model was an improvement in computability. We now have an easily solvable expression for the probability . However, we still have the problem of dependence on the unobservable and . This was solved by Bradley (1952)[[3]](#footnote-3), giving their names to the model of comparative judgment used in modern education[[4]](#footnote-4).

Using the probability expression above, the *Bradley-Terry* model takes in a set of *N* pairwise comparisons on *M* distinct artefacts, and outputs a set of numerical values, , estimating the perceived volume of our given construct for each artefact Notice that we recycle the notation from earlier, saying that is the estimate for artefact *i*. This is a direct analogy whereby the are in fact estimates of the modal discriminals (the peak of the discriminal dispersion determined by the set of discriminal processes).

In summary, the Bradley-Terry model uses a Maximum Likelihood procedure (Rasch, 1960) based on the number of comparisons ‘won’ by each artefact. By comparing the number of comparisons an artefact actually wins with the number we expect it to win, we can iteratively improve our estimate for each artefact.

To this end, we start by computing a raw score for each artefact. For this, we distill our *N* decisions into numerical values by saying when *A* beats *B*, and otherwise. Notice either when *B* beats *A* or when the artefacts *A* and *B* are not compared. We determine the raw score, , as the number of comparisons won by *A*:

This raw score is, in itself, an estimate for each artefact. However, the raw score is not sensitive to the comparison set for each artefact. It is rarely possible to compare each artefact with every other. An artefact with a low volume of the given construct could receive a high raw score by being compared only with other ‘low scoring’ artefacts[[5]](#footnote-5). With this in mind, we want to generate a more nuanced estimate for each artefact.

To this end, we estimate the raw scores by replacing the binary values from with the probability, This gives a new estimate,

We have now returned to an earlier problem, where is a function of the , the very values we are eventually attempting to estimate. However, we now have all the tools in place to determine an iterative expression for the quality of using the Newton-Raphson method as follows:

With each iteration, we improve the estimate for each artefact based on the most recent for each script. All that remains is to determine an initial state for the iterative process. For this, we have a ready-made set of candidates in the raw scores, . In this way, write , where the superscript indicates the number of iterations of the Newton-Raphson method used. Notice that , the current estimate for artefact *i* at the *kth* iteration, is based on the for and where . This process generates an increasingly accurate score for each artefact at each iteration, and likely stabilises after approximately iterations (Pollitt, 2012).

#### A note on generating pairings

Acknowledge new defense of ACJ Kimbell, R. (2021). Examining the reliability of Adaptive Comparative Judgement (ACJ) as an assessment tool in educational settings. *International Journal of Technology and Design Education*, 1–15. <https://doi.org/10.1007/s10798-021-09654-w>)

**EVALUATING THE SCALE**

Having discussed the theoretical models and numerical computations used to estimate the quality of each text, we now turn to mechanisms for evaluating the merits of the scores produced by the Bradley-Terry model. We do so through the lenses of reliability and validity, considering multiple approaches in both cases.

### Reliability

There are three standard measures of reliability prevalent in the literature. Here, we discuss the statistic details of each, alongside potential measurement problems stemming from each.

#### Scale Separation Reliability

In comparative judgment research, Scale Seperation Reliability or SSR is the most commonly reported measure of agreement amongst judges. In literal terms, SSR measures how well the assessment separates the texts. The name SSR takes has origins in Rasch analysis and was first introduced by Bramley (2007) following the observation that several authors were reporting the same measure by different names. Andrich (1978) showed that SSR can be interpreted similarly to Cronbach's alpha, and is thus often interpreted as a measure of internal consistency, with the same thresholds for success as acceptable).

To compute SSR, we first compute a Separation Coefficient, , where is the standard deviation of the estimates and *rmse* is the root mean square of the estimation error[[6]](#footnote-6). This is then converted into Scale Separation Reliability,

While understood as a robust measure of internal consistency, SSR is sensitive to over-estimation based on the size of the data and the type of comparative judgment algorithm used (Jones et al., 2019; Verhavert et al., 2018).

In response to the potentially prohibitive volume of data required to generate reliable scores, researchers have sought to adapt the standard algorithm to generate pairings in which more `information' is generated by each judgment (Pollitt 2012). Adaptive comparative judgment reportedly generates stable scores with fewer judgments than the non-adaptive approach (*ibid).* However, it also artificially inflates SSR, leading to a problematic basis upon which to conduct reliability research (Bramley & Vitello, 2019; Bramley & Wheadon, 2015). With this in mind, we recommend that researchers use only non-adaptive comparative judgment.

SSR increases with the number of judgments, meaning that one gets a higher estimation of reliability simply by collecting more judgments. From a practitioner's perspective, this is arguably an asset, providing an indication of the minimal input necessary to produce a stable output. However, for research purposes, this is problematic particularly when the reliability of a measure is in question. This sensitivity also limits the meaningfulness of comparisons across studies with varying numbers of judgments.

#### Inter-rater reliability

A second, arguably more robust approach to reliability comes from the split-half method introduced by Bisson et al. (2016), referred to as *inter-rater reliability*. To compute this measure, judges are split, post-judging, into two randomly generated groups and scores are calculated for each group using the standard Bradley-Terry model discussed earlier. Reliability is then estimated by computing the Pearson Product-Moment correlation coefficient between the two groups. This procedure is repeated 100 times and the median correlation coefficient yields the measure we refer to as inter-rater reliability. Knowing that reliability increases with data size, this split-half process usually generates an under-estimate of reliability as a result of only using half the decisions in each isolated calculation. As a result, researchers are compelled to collect more data to generate the same conclusions than if they used only SSR.

This stricter measure of reliability is not as sensitive to the number of judgments and, despite the necessity for more data, has gained popularity in recent literature (Ben Davies et al., 2020; Jones et al., 2019; Verhavert et al., 2018). In a meta-analysis of 49 comparative judgment-based studies, Verhavert et al. (2019) demonstrated that this split-half measure is significantly correlated with SSR and can therefore also be meaningfully interpreted as a measure of internal consistency. This suggests that the more resource-intensive approach to reliability may be unnecessary.

Verhavert et al. (2019) also found that in general, one requires only 12 judgments per script to expect to reach an acceptable threshold, With split-half reliability in mind, we recommend that researchers aim to collect 20 judgments per script in exploratory research where comparative judgement is not a proven method in the field. This is more than enough to evaluate SSR and facilitates inter-rater reliability analysis based on approximately 10 judgments per script. Once reliability has been established in a particular field of study, researchers might then consider reducing their requirements to Verhavert et al.’s 12 judgements per script.

#### Judge and script misfits

A third possible approach to reliability is a measure of an item's fit to the model. For every pairwise comparison, it is possible to deduce a measure of `surprise' (Pollitt 2012, p. 164). inherent in that comparison. The degree of surprise, or fit, is quantified by the difference between the expected and observed values. By considering the surprise inherent in decisions made by a given judge, one can produce a measure of the misfit for that judge.

The role and use of misfit in the literature has been inconsistent, raising questions about its value for social science research. We position misfit as related to reliability as it is a measure of the difference between judges and can hence be viewed as a proxy for inter-rater reliability. However, its standard usage, set out by Pollitt (2012) is one regarding quality control and is hence more closely akin to (external) validity. While it is reasonable to evaluate the quality of a given dataset by investigating the number of judges (and scripts) behaving unexpectedly, two problems arise when using this measure to consider excluding data. These stem from the tension between misfit as a measure of reliability or validity.

The first is a recursion problem. After excluding the misfit data, one presumably computes a new model and checks for misfit data again. It is likely that new data will now appear as misfitting. This problem can be solved with pre-registered analysis. However, in doing so, the misfit measure becomes a tool to improve the quality of the data, but loses its power to evaluate reliability and validity.

Similarly, it is unclear what researchers should do with responses identified as misfits, but that do not appear qualitatively unusual. In social science, comparative judgment is often used on the premise that identifying artefacts is difficult in isolation. A researcher can `examine' a misfit artefact and qualitatively consider its place in the dataset, but this subjective approach appears to somewhat undermine the quantitatively driven method.

We argue that misfit is not a productive tool for research purposes and hence do not recommend that researchers report these values. See Davies (2020) for the technical details of misfit calculations, and a more substantive discussion of the associated limitations.

### Validity

In this section, we discuss three common approaches to generating validity evidence in the literature on comparative judgment: *expert testimony, content analysis, and comparative analysis*.

#### Expert testimony

It is common to investigate the validity of comparative judgment scores using expert testimony, as in (Pollitt & Murray, 1993). This testimony usually serves dual purposes within the research design. Most immediately, researchers compare expert testimony to theoretically expected ideas. For example, in their investigation of primary students' portfolio-based tasks in science and technology, Davies et al. (2012) drew positive conclusions regarding the validity of their task based on the theoretically appropriate priorities expressed by the judges. Similarly, Jones et al. (2015) relied primarily on expert testimony collected via two closed-answer questionnaires in evaluating a comparative judgment-based secondary school assessment designed as an alternative to a standard British GCSE assessment. The first questionnaire asked teachers how well their alternative assessment addressed primary curriculum features. The second asked teacher judges which elements of students' responses most influenced their decision-making. In a less structured manner, Jones et al. (2019) considered validity through discussions with the `project advisory panel' (p. 674) of the task design and responses from a pilot study.

#### Content analysis

It is also common to pair expert testimony with qualitative analyses of the responses being judged (as in Hunter & Jones, 2018). In these cases, researchers have evaluated validity based on comparisons between judges' testimony and a coded content analysis of the task responses.

In investigating primary students' free-response explanations of mathematical concepts, Hunter and Jones conducted a content analysis for a sample of six students' responses. They reported symmetry between the comparative judgment-based scores, expert testimony from interviews with teachers, and the qualitative features of the sampled responses. On this basis, they concluded that their comparative judgment-based assessment had demonstrated validity in this case.

Content analyses have also been used in the absence of expert testimony, instead directly comparing content analysis with comparative judgment-based scores using statistical modelling. For example, Jones and Karadeniz (2016) investigated secondary students' conceptual understanding with a series of open-ended questions evaluated using comparative judgment. In evaluating the validity of their measure, they conducted a qualitative analysis of students' responses, coding them for five important traits predetermined from the literature. After conducting a multiple linear regression predicting comparative judgment scores using these five codes (as well as file size and performance on a standard test on fractions), the authors concluded that their comparative judgment-based evaluation had demonstrated acceptable validity.

#### Comparative analysis

The third approach to validity comes from quantitative comparison with theoretically similar measures. Bisson et al. (2016) considered the validity of their comparative judgment-based assessments of students' conceptual understanding by comparing their new measure with outputs from existing validated measures of theoretically similar entities. For their investigation of students' understanding of *p*-values, they benchmarked comparative judgment scores against performance on the existing RPASS-7 test. It is also common to benchmark comparative judgment-based assessments against standard measures of attainment for the population from which participants are recruited. For example, Jones and Alcock (2014) evaluated their assessment of conceptual understanding in introductory real analysis using the summative assessment scores attained at the end of the module on which students were enrolled. Jones et al. (2015) did similarly with scores in evaluating the validity of their secondary school assessment of mathematical problem-solving.

In a similar vein, Jones et al. (2019) performed a randomised control trial using a comparative judgment measure of students' algebra performance alongside a suite of standardised measures of procedural understanding, conceptual understanding and general achievement, as well as writing skills and mathematics anxiety. Similar to Bisson et al. (2016) they considered the correlation between comparative judgment-based scores and their standardised measure of algebra performance. Jones et al. (2019) also considered the capacity of their comparative judgment-based scores to detect the effect of their RCT intervention (known to exist through their algebra measure), expecting to find divergence between their control and intervention groups.

Jones et al. (2019) further pursued this notion of divergence by comparing comparative judgment-based scores of algebra performance with writing skills, finding a moderate correlation. This was explained as a function of the primary school setting in which students' ability to produce coherent sentences was likely related to their ability to respond to the comparative judgment prompt. Nevertheless, the authors drew no explicit conclusions about validity from this evidence. With a similar method, \cite{Jones2016} found that comparative judgment scores correlated more strongly with general mathematics achievement than with reading achievement, and hence that they had found further evidence for the validity of their measure of conceptual understanding.

Jones and Alcock (2014) also considered divergence as a measure of validity, albeit via a different mechanism. By judging their introductory analysis assessment with expert, external non-expert and peer judges, they evaluated the extent to which the scores produced were based on inherently mathematical features. Upon finding a significant difference between models produced by the judging cohorts, these authors concluded their data demonstrated validity as an assessment of mathematics, rather than non-mathematical features upon which the non-experts judges were assumed to have focused.

**DESIGN CONSIDERATIONS**

The past decade has seen growing interest in comparative judgement methods amongst education researchers. In this section, we draw on this corpus in order to identify study design principles for running comparative judgement experiments. We address the common decisions researchers must make to run a comparative judgement experiment including which software to use, what objects are to be judged, who to recruit as judges and how to prepare them, and quantitative decisions such as the number of judgements needed for a given number of objects. We then outline *n* possible analysis templates researchers might choose to adopt in addressing their chosen phenomenon of interest.

**Comparative judgement software**

Software not necessarily needed, e.g. Thurstone’s study and other examples where lists need only be provided.

However if the objects need to be viewed directly rather than signified by a verbal label or other simple symbol then we need software

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Price |  |  |  |  |
| RM Compare | Limited free trial available |  |  |  |  |
| No More Marking | Free to researchers and teachers |  |  |  |  |
| D-PAC |  |  |  |  |  |
| Moodle plug-in | Free to researchers and teachers |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Online platforms available: with an emphasis on No More Marking Ltd.

What objects are suitable and how to generate them

Who is qualified to do the judging and what training they might need

Quantitative decisions: number of judges, judgements, objects

Use language of anchors and baselines rather than grade boundaries/cut scores

**DISCUSSION**

A simple idea, with many potential applications across the social sciences

Common objections: it’s black box; it’s inherently norm-referenced; it’s inefficient

**References**

Andrich, D. (1978). Relationships Between the Thurstone and Rasch Approaches to Item Scaling. *Applied Psychological Measurement*, *2*(3), 451–462. https://doi.org/10.1177/014662167800200319

Bisson, M., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring Conceptual Understanding Using Comparative Judgement. *International Journal of Research in Undergraduate Mathematics Education*, *2*(2), 141–164. https://doi.org/10.1007/s40753-016-0024-3

Bradley, R., & Terry, M. (1952). Rank Analysis of Incomplete Block Designs. *Biometrika*, *39*(3–4), 324–435. https://doi.org/10.1093/biomet/42.3-4.450

Bramley, T. (2007). Paired Comparison Methods. In P. Newton, J. Baird, H. Goldstein, H. Partrick, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 246–300). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/487059/2007-comparability-exam-standards-i-chapter7.pdf

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, *26*(1), 43–58. https://doi.org/10.1080/0969594X.2017.1418734

Bramley, T., & Wheadon, C. (2015). The reliability of Adaptive Comparative Judgment. *AEA-Europe Annual Conference*, *March*, 7–9.

Davies, Ben, Alcock, L., & Jones, I. (2020). Comparative judgement, proof summaries and proof comprehension. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-020-09984-x

Davies, Benjamin. (2020). *Comparative Judgment and Proof*. *January*.

Davies, D., Collier, C., & Howe, A. (2012). Assessing scientific and technological enquiry skills at age 11 using the e-scape system. *International Journal of Technology and Design Education*, *22*(2), 247–263. https://doi.org/10.1007/s10798-011-9191-3

Hunter, J., & Jones, I. (2018). Free-Response Tasks in Primary Mathematics: A Window on Students ’ Thinking. *Making Waves, Opening Spaces: Proceedings of the 41st Annual Conference of the Mathematics Education Research Group of Australasia*, 400–407.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, *45*(3), 662–680. https://doi.org/10.1002/berj.3519

Jones, I., & Karadeniz, I. (2016). An Alternative Approach To Assessing Achievement. *Proceedings of the 2016 40th Conference of the International Group for the Psychology of Mathematics Education*.

Jones, I., Swan, M., & Pollitt, A. (2015). ASSESSING MATHEMATICAL PROBLEM SOLVING USING COMPARATIVE JUDGEMENT. *International Journal of Science and Mathematics Education*, *13*(1), 151–177. https://doi.org/10.1007/s10763-013-9497-6

Luce, R. (1959). *Individual Choice Behavior*. Wiley.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *Performance Testing, Cognition and Assessment. Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, *August*, 74–91.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago Press.

Thurstone, L. (1927). A law of comparative judgement. *Psychological Review*, *34*(4), 273–286.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy and Practice*, *26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment? *Applied Psychological Measurement*, *42*(6), 428–445. https://doi.org/10.1177/0146621617748321

Other potential areas:

The economics of tourism, e.g. Boto-García, Mariel, Pino & Alvarez (2020). “*To avoid cognitive burdens, respondents are presented with three alternatives per choice task plus a ‘none of them’ option. This is done to avoid forcing them to choose any of the alternatives if none of them are attractive enough. This is standard practice in DCE in general (Oehlmann et al., 2017) and in the context of holiday choice in particular (Bimonte et al., 2016; Grigolon et al., 2012; Huybers, 2003) to ensure realism.*”

Choice experiments: e.g. “marginal values for the attributes of environmental assets, such as forests and rivers” (Hanley, Wright & Adamowicz, 1998; p. 413)

Examples: Thurstone’s applications, sports example (baseball) education applications; (D&T, maths, English); monitoring standards (Ofqual type usages); generating standards (Heldsinger et al; NMM services); measurement in research (RCT papers);

Beauty in philosophy: Tuya’s work; philosophy abstracts (Matthew’s idea)

Standards in HE <https://www.advance-he.ac.uk/degree-standards-project/calibration-academic-standards>

judging REF environment statements

1. Prior to Thurstone's work, the discriminal process had proven a major barrier to psychologists' attempts to develop meaningful measurements of phenomena subject to high variance and had steered many away from the measurement of subjective phenomena altogether Bramley (2007). [↑](#footnote-ref-1)
2. This can be proven by considering the variance associated with and [↑](#footnote-ref-2)
3. Bradley and Terry (1952)'s original model was more general than Thurstone's, based on a set of less stringent assumptions. In particular, only Thurstone (1927)assumed an equivalence across discriminal dispersion. An in-depth discussion of the consequences of this assumption can be found in Bramley (2007), along with a justification for the numerical and theoretical equivalence between the two. [↑](#footnote-ref-3)
4. Luce (1959) presented very similar work and on occasion, this model is referred to as the `Bradley-Terry-Luce' model (E.g. Verhavert et al., 2018). In line with the majority of the comparative judgment literature, we refer only to the Bradley-Terry model from here on. [↑](#footnote-ref-4)
5. In principle, one could imagine a dataset with every artefact compared with every other exactly once and this problem disappears. However, such a dataset is impractical to generate. [↑](#footnote-ref-5)
6. The estimation error for each is computed using the inverse of Fisher's information Matrix, equivalent to the covariance matrix. [↑](#footnote-ref-6)