European Journal of Operational Research 228 (2013) 148–157

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

# Comparing the validity of numerical judgements elicited by direct rating and point allocation: Insights from objectively verifiable perceptual tasks

Paul A. Bottomley *, John R. Doyle [1]

*Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff CF10 3EU, United Kingdom*

ARTICLE INFO

*Article history:*
Received 22 February 2011
Accepted 3 January 2013
Available online 4 February 2013

*Keywords:*
Decision analysis
Methods of value elicitation
Direct rating
Point allocation
Validity
Objective tasks

ABSTRACT

Two popular methods for assigning numerical values to a set of to-be-judged objects in order to capture their relative standing are Direct Rating (DR) and Point Allocation (PA). People using PA distribute a fixed sum of 100 points among the objects, while people using DR rate each object on a fixed scale, typically 0–10, later rescaled to sum to 100. Prior research shows that these methods exhibit distinct profiles when values are ranked from largest to smallest, with DR being more test–retest reliable. But which method best translates people's inner judgments into outer numerical values (is more valid)? Instead of examining subjective or abstract stimuli, we use objectively verifiable perceptual tasks, namely judgments of line length presented using bar charts. We show that (i) DR is more inter-rater reliable than PA; (ii) DR is more accurate than PA at the individual level; (iii) but there is no difference in accuracy when individual judgments are combined to form group-level estimates; and (iv) DR judgments were improved by using prior knowledge of method bias, whereas PA judgments were not.



## 1. Introduction

A great deal of management research is concerned with making numerical judgments to capture the relative standing of a set of objects. Psychologists are concerned with things such as people's goals and values. In organizational behaviour, managers are interested in how employees should (and do) allocate their time among different activities, as epitomized by the classic time-and-motion study. In marketing, segmentation endeavours to uncover groups of likeminded people who are looking for similar benefits in a product. In new product development, the judgments of team members are often aggregated (averaged) before deciding which ideas warrant further screening. This paper concerns all such numerical judgments. To avoid any ambiguity among the different vocabularies used in these applications, we adopt Doyle's (1999) terminology and use the word object to refer to the super-ordinate category that includes attributes and alternatives, be they goals, activities, benefits, ideas, or anything else that is numerically judged. Such judgments return values, which is our overall term for weights, ratings and probabilities.[2]

Not surprisingly, many methods of elicitation have been proposed for obtaining these numerical values. Unfortunately, research assessing attribute importance, coupled with studies on the psychology of survey response, suggests that the choice of method is not arbitrary (for comprehensive reviews, see von Winterfeldt and Edwards, 1986; Weber and Borcherding, 1993; Tourangeau et al., 2000; Pöyhönen et al., 2001; Slovic et al., 2007; Morton and Fasolo, 2009). Direct Rating (DR) and Point Allocation (PA), the focus of this article, are two such popular methods that violate the principle of procedural invariance (Arrow, 1982; Slovic, 1995) because these notionally equivalent methods give rise to different numerical values.

People using PA distribute a fixed sum of 100 points among the objects, while people using DR rate each object on a fixed scale, typically 0–10, later rescaled to sum to 100. At first sight, these methods appear to be minor variants of one another, involving a simple arithmetic translation, but they do exhibit distinct rank-to-value profiles when the objects are ordered from largest to smallest (Doyle et al., 1997). People using PA gave nearly 50% more value to that object ranked most important as those using DR and conversely 50% less value to that object ranked least important. Van Ittersum et al. (2007) interpreted this differential response as signalling poor convergent validity. Likewise in tests of predictive validity, where the values (weights) are combined with information on hypothetical alternatives to explain consumer choice (Srivastava et al., 1995; Bottomley, Doyle and Green's, 2000), the level of agreement has been modest. Nevertheless, values elicited by DR are more test–retest reliable than those of PA when measured 1 week apart (Bottomley et al., 2000). So,

* Corresponding author. Tel.: +44 2920 875609; fax: +44 2920 8726473.
*E-mail addresses:* bottomleypa@cardiff.ac.uk (P.A. Bottomley), doylejr@cardiff.ac.uk (J.R. Doyle).
[1] Tel.: +44 2920 875695; fax: +44 2920 8726473.

[2] In decision analysis, multi-attribute models simultaneously integrate information from two sets of numerical judgments, one measuring attribute importance and the other alternative values. While importance weights are expressed in relative terms, alternative values are expressed in absolute terms – Product A can outperform Product B on all criteria – thereby making them beyond our remit.

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.01.005

collectively these studies suggest that DR is superior to PA, but the evidence is far from overwhelming.

Almost all previous research comparing PA and DR has relied exclusively on people making numerical judgments of privately held opinions. For instance, in Doyle et al. (1997, Exp. 2), people were asked to judge the same set of objects twice, once using PA and once using DR. Since the two evaluations differed systematically, it is clear that one or both methods of elicitation must be distorting people's privately held opinions. If we could peer inside the metaphorical "black-box" that is the decision-maker's mind, we could extract each person's true, undistorted values of the objects. Armed with such knowledge, we might then assess the ability of PA and DR to faithfully reproduce this information. However, since we do not have privileged access to people's opinions (they remain in a black-box), the next best thing is to examine problems with objectively verifiable answers, which is the approach we take in this article.

But what are suitable problems? Although the fuel consumption of automobiles or the market share of brands is verifiable, it may not be widely known, thus introducing extraneous noise into any comparison of methods. To avoid relying on general knowledge or tests of memory, we will use visual stimuli. But recent research in marketing on the consumption of foods has suggested that judgments of area (e.g., size of pizza slices) and volume (e.g., capacity of drinking glasses) are prone to distortion (Wansink and Van Ittersum, 2003). Studies on the graphical representation of information reach similar conclusions with estimates of length being more accurate than area and volume (Cleveland and McGill, 1984; Lurie and Mason, 2007). Accordingly, we focus on judgments of line length, presented in the form of bar charts. Steven's (1975) work in psychophysics has also shown that length judgments have the desirable quality of proportionality; a 12 inch line is actually seen as being twice as long as a 6 inch line. Thus, by using lines, we can ensure that people perceive the stimuli "truly" and deviations in their numerical judgments can be attributed to the method itself and not to errors in perception.

This research makes several contributions. First, we contribute to those studies exploring the connection between people's judgmental perceptions, methods of elicitation and the issue of procedural invariance. Second, we consider objective tasks that have palpably verifiable answers instead of subjective tasks. This innovation allows us to introduce new measures that directly rather than indirectly assess elicitation method performance. Third, analyses are performed at both the individual and aggregate level because decisions may focus on pooled estimates or those of each respondent. By addressing the heterogeneity of individuals rather than simply averaging responses, we can explore the potential impact of aggregation bias on our findings. Fourth, if people consistently under or overestimate the value attached to particular rank-ordered objects, we also can determine how useful knowledge of "method bias" is for adjusting the numerical judgments elicited by a different, independent set of respondents (cross-validation).

But, before beginning in earnest, and to dispel any confusion about what we aim to achieve in this paper, it may be useful to state briefly how this paper differs from more frequently encountered research in value elicitation which either seeks to augment value-elicitation from incomplete data (e.g., Barron and Barrett, 1996; Jessop, 2004; Bous et al., 2010; Conde and del la Paz Rivera Pérez, 2010; Kadziński et al., 2012), possibly in a group (Yeh and Chang, 2009) or even, a population context (Musal et al., 2012); or resolves error-prone human judgment, such as the intransitivity of pairwise preferences (Siraj et al., 2012). Despite their differences, these studies are all examples of research that aims to improve fallible human judgments by presenting new mathematical models that translate linguistic, or ordinal, or fuzzy, or inconsistent judgments into numbers, in the best engineering traditions of cleaning up a weak or noisy signal.

By contrast, this paper makes no attempt to present a new mathematical model. Acknowledging that PA and DR are methods that will probably never go away, and will continue to be used widely in preference to the more sophisticated methods mentioned above, this paper takes DR and PA as givens, in all their naivety. The paper sets itself the task of deciding which method should be preferred. The art of this paper is to devise a test that distinguishes in this way between PA and DR; and to collect and analyse data with sufficient rigour that the conclusions which flow from it can be accepted with confidence. The data and analyses are therefore not perfunctory illustrations of new mathematical models, but are the heart of the enterprise. Similarly, Linares (2009) and Ishizaka et al. (2011) used experimental data and statistical analyses to examine aspects of the Analytic Hierarchy Process – although neither used an objective criterion, as we do, as a 'gold standard' against which to compare methods, instead drawing their conclusions based on the convergence between methods used to elicit subjective evaluations.

## 2. Theoretical background

Doyle et al. (1997) compared the properties of Direct Rating (DR) and Point Allocation (PA) across a variety of decision scenarios, including the skills that potential employees thought a retailer might require of managerial recruits. Having sorted each person's attribute weights (rescaled to sum to 100) from most to least important, the elicitation methods were found to exhibit distinct patterns of weights in the rank positions (rank-to-value profiles). In particular, the slope of the rank-to-value profile was steeper for PA than DR. Its curvature was also more pronounced (sagging rather than billowing). In fact, the profile was distinctly curvilinear for PA but linear with rank for DR. People gave nearly 50% more weight to their most important attribute using PA than DR, and similarly 50% less weight to their least important one.

An obvious next question is whether the form of the rank-to-value profile is unique to problems where the objects to be judged comprise attributes to be weighted? Across a variety of tasks, Doyle (1999) showed that it also applies to ratings, and more subtly, problems naturally conceived of as fixed-sum, regardless of elicitation method. Applying Gardner's (1983) theory of seven multiple intelligences, students' values of both themselves and a friend were found to be curvilinear with rank using PA, but linear with rank using DR (Exp. 1). Because people differed greatly in terms of qualities possessed (low concordance), this lack of agreement helps to rule out the possibility that these rank-to-value profiles were an artefact of the objects valued and not their rank positions. Doyle also showed that budgeting and other top-down methods for allocating finite resources (e.g., time, prize monies) implicitly imposes a fixed-sum constraint on the problem, just like PA does on the values. Making this constraint explicit, the values elicited by both PA and DR exhibited the distinct curvilinear rank-to-value profile, though more so with PA than DR.

Convergent validity has typically been examined using aggregate-level analyses, where the mean values (not rank ordered) of people using one method are correlated with the corresponding mean values of those using another method (e.g., Zhu and Anderson, 1991; Schori, 1995). Jaccard et al. (1986) compared six elicitation methods, including DR and a variant of PA, but found the evidence for convergent validity to be "far from impressive". They speculated that attribute importance might be a multidimensional construct, with different methods tapping different aspects. Van Ittersum et al.'s (2007) meta-analytic study develops this idea. They classified ten popular methods according to one