2016

A respondent friendly method of ranking long lists

James E. Heyman
jeheyman@stthomas.edu

John Sailors
john.sailors@scranton.edu

A Respondent-friendly Method of Ranking Long Lists

Abstract

This article illustrates a consumer-friendly approach to preference elicitation over large choice sets that overcomes limitations of rating, full-list ranking, conjoint, and choice-based approaches. This approach, HL*m*, requires respondents to identify the top and bottom *m* items from an overall list. Across respondents, the number of times an item appears in participants' L (low) list is subtracted from the number of times it appears in participants' H (high) list. These net scores are then used to order the total list. We illustrate the approach in three experiments, demonstrating that it compares favorably to familiar methods while being much less demanding on survey participants. Experiment 1 had participants alphabetize words, suggesting the HL*m* method is easier than full ranking but less accurate if *m* does not increase with increases in list length. The objective of experiment 2 was to order U.S. states by populations. In this domain, where knowledge was imperfect, HL*m* outperformed full ranking. Experiment 3 involved eliciting respondents' personal tastes for fruit. HL*m* resulted in a final ranking that correlated highly with max-diff scaling. We argue that HL*m* is a viable method for obtaining aggregate order of preferences across large numbers of alternatives.

Introduction

Suppose a marketing manager wants to know consumer preferences for the various ice cream flavors her company produces. Inferring preferences from sales data is problematic because not all flavors have the same distribution intensity, shelf space in the stores that carry them, or promotional support. Modeling preferences by accounting for flavor dissimilarities is complicated and time consuming, and so the manager turns to surveying consumers. A contemporary approach would be to use adaptive-choice-based conjoint or maximum difference scaling, but these require sophisticated software and data analysis expertise. Rating each flavor is straightforward, but typically results in little variance across items. Ranking is also straightforward and offers greater discriminatory power. However, ranking is only recommended for very short lists of items (Sudman & Bradburn 1982, p. 149) and flavors of ice cream (and perfume scents, automobile colors, and other product category attributes) are numerous. Ben & Jerry's offers 58 flavors of ice cream, though this number does not include frozen yogurt, and the company introduces new flavors continually. This situation generalizes to online companies that offer larger sets of items even when fewer people buy each individual product {Brynjolfsson, 2003 #210}. These "long tail" marketplaces blur the size distinction between choice sets, consideration sets, and awareness sets. In these environments, short lists are important, not because they represent consumers' decisions but merely because of ranking items is so cognitively taxing. Ranking's cognitive demands cause people to become sloppy when engaged in the task. For example, when we had 98 students alphabetize 30 words, only 32% did it correctly. We propose an alternative to full ranking tasks that overcomes the negative aspects of full ranking while producing equivalent results.

Ranking tasks are common in marketing research, used to study diverse topics such as consumer choice (Caparros, Oviedo, & Campos 2008), brand beliefs (Barnard & Ehrenberg 1990), attribute assessment (Vanleeuwen & Mandabach 2002), and customer satisfaction (Durkin 2007). The principle benefit of ranking is that it forces respondents to delineate among items being measured (Klein, Dülmer, Ohr, Quandt, & Rosar 2004). This method accords with the realities of a market in which a consumer eventually chooses one product over others—the essence of ranking (Kamakura & Mazzon 1991). Other benefits are that it forces respondents to use a common scale to assess alternatives (Krosnick & Alwin 1988; Vanleeuwen & Mandabach 2002) and accommodates inconsistent preferences due to variety-seeking (Buchanan, Givon, & Goldman 1987). These benefits, however, come at a cost. Foremost, the ranking quickly becomes untenably difficult as the number of items grows (Alwin & Krosnick 1985).

The problem with ranking

At the core of respondents' difficulty with ranking tasks is that it requires cognitive effort that grows non-linearly as a list of items grows. Respondents find ranking difficult, and the resulting mental fatigue reduces the quality of data (Beatty, Martin, Yoon, & Kahle 1996). Higher cognitive effort also leads to increased costs and difficulty of administering surveys (Munson & McIntyre 1979; McCarty & Shrum 1997) since researchers use labor-intensive methods such as card sorts (Barnard & Ehrenberg 1990) or a computerized equivalent of drag-and-drop. These difficulties preclude using methods such as telephone (Ovadia 2004) or paper-and-pencil surveys. The effect is exacerbated as the list of items gets long (Feather 1973). These costs are inherent to ranking and can be formalized in terms of ranking's inherent algorithmic complexity (Edmonds 2008).

The order of magnitude of the work required to sort n items can be approximated using logic. Ranking n items consists of taking each item and placing it in one of k ordered bins. In the case of full ranking, there are as many bins as there are items (n=k). For the first bin, there are n items from which to choose. The next step requires looking at the remaining n-1 items to choose one for the second bin. The third bin requires reviewing n-2 items, the fourth n-3, etc. By the nth bin, one has looked at $n + (n\text{-}1) + (n\text{-}2) + \ldots + (n\text{-}n\text{+}2) + 1$ items, the sum of the numbers 1 to n, $(n^2+n)/2$. This value is dominated by the n^2 term, the number of items multiplied by the number of bins (Edmonds 2008). The implication for full ranking is that as the number of items doubles, the effort required quadruples; as lists grow, full ranking gets very difficult quickly.

The solution of partial ranking

An alternative that addresses full ranking's shortcomings is partial ranking, a process akin to full ranking except there are only k bins where $k < n$. Thus, partial ranking assigns items to ordered bins that are not limited to a single member. If a bin has multiple items, they are not sorted within the bin. Partial ranking therefore allows ties accommodating similarity and differences among items. Since $k < n$, partial ranking is less cumbersome, and the effort required with this method can again be approximated by multiplying the number of bins (k) by the number of items (n), resulting in kn. The effort required therefore is fixed in k and linear in n, as opposed to being quadratic in n as with full ranking, and is proportionally easier than full ranking as k decreases.

Partial ranking's effectiveness and ease of use appears in studies that position individuals within social groups (Cillessen & Bukowski 2000). A common method has respondents identify the top and bottom m of a population, equivalent to a partial ranking system with k=3 bins, with two bins of size m and one of size n-2m. These data can be analyzed in a variety of ways (Peery 1979; Coie, Dodge, & Coppotelli 1982; Newcomb & Bukowski 1983). One computationally

efficient approach calculates the difference between the number of top and bottom nominations an individual receives from all respondents. We call this High-Low ranking, and hereafter refer to such scores as HL*m*, where *m* is the number of nominations elicited at each end of the ranking. It is important to note that HL*m* is not just a procedural nicety but, particularly within marketing, represents a specific conceptualization of the consumer decision process. A consumer searching Amazon.com for a Seiko day date wristwatch finds 31 alternatives. One subset consists of watches that closely match what the consumer envisioned; one subset has watches that he/she would never buy, and the third and largest subset has watches that he/she wasn't looking for but have positive attributes that he/she hadn't thought of. This is a fundamentally different logic to, say, the as-if psychology of conjoint's pairwise comparison or a "Top m" method such as TURF (Total Unduplicated Reach and Frequency) analysis that focuses on people's choice set rather than their consideration set.

HL*m*'s ease of use makes it attractive in domains in which respondents experience difficulty making numerous comparisons such as assessing bullies (Henry 2006), giftedness (Gagne 1998), romantic partners (Simon, Aikins, & Prinstein 2008), perceived athletic ability (Dunn, Dunn, & Bayduza 2007), leadership (Charbonneau & Nicol 2002), and job performance (Schwarzwald, Koslowsky, & Mager-Bibi 1999). However, it has not been applied to and studied within marketing research. Since peer nominations are still about ranking opinions, they apply as easily to products and consumer preferences as to people's traits. For this method to be useful to marketing research, it must be both accurate and easier than full ranking. We present three studies that demonstrate the ease and accuracy of HL*m* in three cases: when there is a true answer known to the participants, when there is a true answer largely unknown to the participants, and finally the most relevant case of eliciting personal opinions.

Study 1: Alphabetizing Words

We begin by illustrating the ease of the HL*m* method in comparison to full ranking for large *n* in the context of a simple cognitive task: alphabetizing words. Although clearly an impractical way of writing a dictionary, starting with a task with a known true answer helps demonstrates how HL*m,* works and has the benefit of applying implicit social pressure on participants to apply themselves. The conditions provided a test of efficiency between HL*m* and full ranking, while the objectivity of alphabetizing words allowed measurement of effectiveness.

H1: HL*m* tasks require less time for respondents to complete than full ranking of the same item list.

H2: HL*m* tasks will result in comparable accuracy compared to full ranking of the same item list.

Participants

Five-hundred twenty-two undergraduate university students participated for partial course credit in a marketing class and performed these tasks during a one-hour lab session.

Design

We deployed the study using Qualtrics, which allowed us to capture the time required to complete the tasks. Mode of Alphabetizing (i.e., full ranking versus HL5) was a between-subjects factor, and Word List Length was a within-subjects factor. Three-hundred one students were assigned to the full-ranking mode in which they sorted word lists alphabetically using a drag-and-drop task. Two-hundred twenty-one students were assigned to the HL5 Mode in which they identified the first five and last five words, alphabetically, from the word lists. Participants in both modes practiced with a list of 10 words, and then were given lists of words that increased in length by 15, 20, 25, 30, 35, and 40 items. Each set was drawn at random from a master list of 60 two-syllable words, ranging from four to seven letters. For full ranking, we formed the

aggregate list by averaging how participants ranked each word. For HL5, we summed the number of times a word appeared in respondents' top five bins and subtracted the number of times it appeared in the bottom five bins[1]. The words were then sorted according to these scores.

Results

We tested the amount of time it took for people to complete the task, a factor that reflects respondent motivation and, particularly in a computer-mediated environment, is an accurate proxy for cognitive effort. The time required for the task exhibited the hypothesized pattern. Shown in the Figure 1, respondents in the full ranking condition took 74 seconds to alphabetize the 15-word list and needed 233 seconds to alphabetize the 40-word list. A repeated-measures ANOVA (with Box's conservative correction) indicated significant Mode X Word List Length, $F(1, 515)=70.34$, $p<.001$, driven by differences across list length within the full-ranking condition, $F(1,295)=132.1$, $p<.001$, supporting H1.

TAKE IN FIGURE 1

We calculated each method's accuracy with Kendall's τ_a as the primary measure. Although less common than other non-parametric measures such as Spearman's rank correlation, it offers properties appropriate for this study. τ_a measures the number of concordant pairs between two rank orders. This is relevant since the factual nature of the ranking task provides a normative benchmark (Cliff 1996). It also captures the sorting process that is at the core of ranking better (Edmonds 2008), and offers the benefit of intuitive interpretation, especially in

[1] The net score's range is dependent on the size of the sample, not the list. For example, if 200 people use HL*m* to complete a ranking task, the Hlm scores would range from -200 to +200. The ranks, of course, would follow the size of the list. Additional manipulations of the net score are possible, should the researcher desire, say, for comparability to other samples. This could be accomplished by multiplying each items net score by 100/N; any such transformation constitutes an affine shift and thus would have no impact on the resulting ranks.

comparison to Spearman's ρ (Wilkie 1980). The results are robust whether analyzed with τ_a or ρ.

Perfect knowledge of the domain did not translate into perfect accuracy. In the full ranking condition, Kendall's τ_a ranged from 0.94 for the 15-word list to 0.98 for the 40-word list. Comparing the ranks derived from the HL5 method with the correct word order, τ_a ranged from a low of 0.76 for the 40-word list to a high of 0.92 for the 20-word list. Kendall's τ_a for HL5 tracked well with the tau for full ranking for the 15- and 20-word lists, but then declined. Figure 2 compares τ_a for the two Mode conditions across list lengths. Thus H2 was supported for relatively shorter lists but not for the longer lists.

TAKE IN FIGURE 2

Discussion

Response times across the two ranking methods illustrate that full ranking becomes increasingly difficult as the set to be ranked increases, but this was not the case with HL*5*. As the lists became longer, respondents in the full-ranking condition needed an average of 5.2 seconds longer to respond per word, whereas in the HL5 condition, each additional word added only .24 seconds. Unfortunately, as the number of words increased, HL*m*'s accuracy decreased. This arises from keeping m=5 when the list becomes long, and the uniformly high ability regarding the task (i.e., alphabetizing). Maximizing the amount of information from HL*m* requires that m should increase to one-third of the list so that the three bins are the same size (Michalowicz, Nichols, & Bucholtz 2013). This condition was met when N=15, decaying as the length grew to 40 items; the decrease in information was proportional to the decrease in accuracy.

This accuracy decay is due to a lack of variance in individual HL*m* responses, which led to a restricted number of aggregate categories and a correspondingly low tau when compared to true rankings, rather than an innate shortcoming of the method. At the extreme, with perfect inter-respondent consistency, the HL*m* task results in 3 aggregate categories (i.e., m objects tied each in the high and low categories, and n-2m in the middle), which compares poorly with any list whose true ranking has more than 3 categories. The HL*m* method does best when there is variability among respondents. In this study the lack of variability is inherent to the task but HLm will have the same problem in domains with a small number of dominant choices as opposed to the more uniform distribution found in long tails. This variability might be due to differences in knowledge or opinion, and is crucial to the HL*m* aggregate data being greater than the sum of its inputs. The second study involves a different ranking task, one for which respondent knowledge is imperfect.

Study 2: Ranking State Populations

We again use another objective ranking task, but one in which there is significant variance in respondents' knowledge—ranking the states of the United States by population. This allowed us to explore how consensus ranking using HL*m* compares with full ranking regarding accuracy when knowledge is imperfect and varies across individuals, a condition more common to many marketing research tasks than alphabetizing words. This study also varied the size of the HL bins used in the HL*m* conditions. Bins of size n/3 are optimal, so larger HL bins should result in better performance of HL*m* (Michalowicz, Nichols, & Bucholtz 2013). Contrary to the first study, we kept *n* fixed at 25, but participants used HL3 on one set of states and HL5 on the remaining 25.

H3: The HL5 task will result in an aggregate ranking comparable to full ranking.

H4: The HL5 task will result in a more accurate aggregate ranking than will the HL3 task.

Participants

Twenty-seven undergraduate university students participated for partial course credit in an introductory marketing class.

Design

Sorting method was a within-subjects factor. The study consisted of three tasks in which participants sorted sets of 25 states by population. During one task, participants identified the three states they perceived had the largest populations, and the three with the smallest. During the second task, participants identified from a list of the other 25 states the five states they perceived had the largest and smallest populations, respectively. During the final task, participants fully ranked a list of 25 states, with members drawn from each of the two previous lists. To motivate respondents, we offered accuracy-based financial incentives tied to performance on the full ranking task, the most onerous of the tasks assigned. The most accurate participant won $30, the second $20, and third $10. Participants were also told that everyone who ranked all 25 states correctly would receive $100. All three tasks were completed on paper forms, and participants were provided with pencils and erasers.

Results

Each state's HL3 and HL5 rankings were calculated by subtracting its total number of low votes from its high votes. These totals were then ranked in descending order and compared to the states' correct rankings. The full ranking results were based on averaging each state's rankings across the participants.

TAKE IN FIGURE 3

In contrast to information content expectations, results from the HL5 method (τ_a (23)=.76, p<.001) were not different from those of full ranking (t(23)=.62, p=.62), supporting H3. Results from the HL3 method (τ_a (23)=.71, p<.001) were commensurate with those of HL5, but again not different from full ranking (t(23)=.13, p=.90); therefore, H4 was not supported.

Table 1 has all three rankings compared to the correct ordering and offers another way of comparing the results.

TAKE IN TABLE 1

Our eventual goal (Study 3) is to apply the HL*m* method to questions that have no pre-knowable correct answer. But, in the absence of measurable accuracy, how does one measure the validity of a ranking method? One way to do this is to see how the various lists correlate with each other. This is largely independent of accuracy because, outside of very high values of Kendall's Tau, two lists can be similarly accurate and yet not be highly correlated to each other. As can be seen in Table 1, all three ranking methods generate highly correlated lists: the correlation of HL3 and HL5 is .95, between HL3 and Full Ranking is .93, and between HL5 and Full Ranking is .93.

Discussion

Despite being known as a difficult task for respondents to complete, ranking remains popular for some researchers and topics for eliciting people's attitudes, opinions, and preferences. One challenge of comparing survey methods is lack of a benchmark against which to measure various techniques. We avoided this problem by using a task with a known correct answer, which allowed us to focus on the techniques' accuracies. The major implication of these results is that partial ranking methods perform as well as full ranking. With regard to obtaining a rank ordering of 25 states by population, both HL3 and HL5 were as accurate as full ranking.

For any individual respondent, HL*m* results in the three categories described earlier. In aggregate, however, and given variance in individual responses, HL*m* results in more than three categories. In the current study, the number of ordered groups of states in aggregate ranks was as follows. With perfect knowledge, full ranking of the 25 states by population results in 25 ordered "groups" of states after aggregation of individual data. With respondents' imperfect knowledge, the aggregated full ranking resulted in the 25 states being sorted into 12 ordered groups of states, HL3 resulted in 16, and HL5 in 19. Intuitively, the fewer the ordered groups, the less information the consensus ranking yields, (Shannon 1948). Therefore, HL3 and HL5 produced more information than full ranking did while requiring less effort from the participants. .

Combining results from Studies 1 and 2, we argue that variability in response, whether due to differences in knowledge, is necessary for the HL*m* methods to result in an aggregate rank that performs as well as full ranking. In the next study, we turn to the more practical research situation; applying HL*m* to differences of opinion.

Study 3: Fruit Preferences

In the first two experiments, we illustrate the use and advantages of the HL*m* method during tasks with objectively correct answers—alphabetizing words and ordering states by population. In this final study, we compare performance of subjective preferences between HL*m* and maximum-difference scaling (MaxDiff), a common method of eliciting importance weights in applied marketing research (Louviere 1991; Finn & Louviere 1992; Sawtooth Software 2013).

H5: HL*m* will produce results comparable to MaxDiff scaling.

Since the domain concerns subjective preferences as opposed to objective facts, this study aligns with the domains of primary interest to marketing researchers. Another feature that distinguishes this study from the first two is use of a non-student sample. In many domains,

including those we use, there is no reason to believe student respondents are not representative of a population, but non-student participants provide a higher degree of generalizability.

Participants

We recruited participants from Amazon's Mechanical Turk (MTurk) service. MTurk is an example of cloud sourcing or distributed computing. The idea is to harness the power of distributed human intelligence by asking individuals, called workers, worldwide to engage in simple tasks that are posted through the MTurk site and for which they receive nominal compensation (in this case, participants were paid $.10 each). MTurk workers report wanting "something fun to do" to be more of a motivation to their participation than "wanting to make money," though "making money while doing something fun" was second. The population of MTurk workers resembles the population, albeit slightly younger, more female, with lower incomes, and from smaller families than the Internet population (Ipeirotis 2008). Two-hundred eighty workers participated in the study.

Design

All workers completed a task that elicited their preferences for fruits from a list of 25 fruits. Some subjects accomplished this using an HL4 task, similar to those described above. The instructions in this case were for subjects to examine the list of 25 fruits and identify their "4 Most Favorite" and "4 Least Favorite" fruits. They were not required to do any type of sorting or ranking within their favorite and least favorite selections, nor among the remaining, unselected fruits. Others accomplished the task using MaxDiff, during which respondents repeatedly identified the most and least desirable product or feature from a carefully constructed subset of the total set under investigation. The name derives from the fact that for each subset, respondents identify a pair of items with the maximum difference between them regarding preference or

importance, or whatever characteristics are of interest to a researcher. Hierarchical Bayesian estimation methods are used on resulting data to calculate individual-level utility functions, which can be subsequently used during other analyses. One simple subsequent use, the one we employ, is to calculate average utility for the items. Participants picked their most and least liked fruit from a list of five fruits drawn from the longer list of 25 fruits, repeated 15 times per respondent.

Results

The Table presents results from both procedures. Since MaxDiff is a common way of deriving preferences, we sorted the fruits by their MaxDiff revealed preferences. Also presented are the preference rank for each method, average utility for each fruit according to MaxDiff (i.e., utilities scaled to sum to 100 within respondent), preference rank according to the HL4 method, and HL4 net score.

TAKE IN TABLE

Visual inspection of the Table suggests the two methods tracked each other well. The MaxDiff (i.e., rescaled) utilities and the HL4 results correlated by .89, thus supporting H5. This produced ranks that also tracked well, differing by 2.7, on average. One exception occurred with cherries, which were ranked much higher under MaxDiff in comparison to HL4 (3rd versus 9th)[2]. The methods agreed on the most-liked fruit (i.e., strawberries), the least liked (i.e., dates), nine of the top ten, and eight of the bottom ten.

Discussion

[2] We have no insight as to why this difference occurred and can thus offer no explanation for it. We suspect it is simply an idiosyncratic difference between results obtained at that particular time.

MaxDiff and related conjoint methods are the most popular ways of establishing preferences. However, they involve complex designs and very complex, computationally intense estimation methods. Study 3 demonstrates that the simpler HL*m* method produces aggregate results that align well with these advanced methods. We are not proposing that HL4 replace MaxDiff or other choice methods. Those methods typically produce individual-level utilities that can then be used for market segmentation and other purposes. HL*m* produces only a market-level picture that cannot be used for market segmentation. However, when aggregate analysis is required, HL4 provides a good solution that despite its simplicity is robust in comparison to advanced methods.

General Discussion

We introduce HL*m*, an aggregated partial ranking, as an alternative to full ranking, particularly for cases involving large choice sets. Results from the first study suggest partial ranking is easier than full ranking. However, during the task in which respondents had near-perfect knowledge, aggregate performance from the HL*m* method was inferior to aggregated performance from full ranking, especially if *m* did not increase as set length increased. Study 2 explores the idea that the uniform level of ability in the first study led to poor quality of results in comparison to known values. In a domain in which individual performance varies, Study 2 suggests respondents using HL*m* outperform as a group what they were able to do as individuals. Study 3 compares the HL*m* method using subjective knowledge against more advanced, complicated, and computationally intense methods popular among applied and academic marketing researchers. Greater ease in comparison to full ranking and its high performance under appropriate conditions makes HL*m* a good candidate for marketing researchers interested in assessing market- or segment-level ranks. However, the method generates data at the individual level, which are sparse and difficult to evaluate with traditional parametric analyses. For example, suppose a

researcher needs to identify segments within respondents. It is unobvious how data from HL*m* cluster without yielding trivial results of clusters of people who have the same three resultant categories. For researchers conducting market-level analyses in domains with many choices, HL*m* is an efficient method of eliciting preferences.

HL*m* is offered as a substitute for full ranking when the desire is, nonetheless, to obtain a fully ordered list of items. This is not an uncommon need in marketing research, as we described in the introduction. Flavors of foods, preferences for new brand names, long tail markets: the applications within marketing are myriad. One thing to keep in mind is that the HL*m* answers the same question as does full ranking; it does not answer other questions such as which combination of list items provides the most complete market coverage (e.g. TURF analysis).

References

Alwin, D.F. & Krosnick, J.A. (1985) The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, **49**, pp. 535–552.

Barnard, N.R. & Ehrenberg, A.S.C. (1990) Robust measures of consumer brand beliefs. *Journal of Marketing Research*, **27**, 4, pp. 477–484.

Beatty, S.E., Martin, W.K., Yoon, M.H. & Kahle, L.R. (1996) *Improving Value Rankings: An Assessment of the Rank-then-rate Approach.* Paper presented at the 1996 Developments in Marketing Science, Miami, FL.

Brynjolfsson, E., et al. (2003). "Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers." Management Science **49**(11): 1580-1596.

Buchanan, B., Givon, M. & Goldman, A. (1987) Measurement of discrimination ability in taste tests: An empirical investigation. *Journal of Marketing Research*, **24**, 2, pp. 154–163.

Caparros, A., Oviedo, J.L. & Campos, P. (2008) Would you choose your preferred option? Comparing choice and recoded ranking experiments. *American Journal of Agricultural Economics*, **90**, 3, pp. 843–857.

Charbonneau, D. & Nicol, A.A.M. (2002) Emotional intelligence and leadership in adolescents. *Personality and Individual Differences*, **33**, 7, pp. 1101–1113.

Cillessen, A.H.N. & Bukowski, W.M. (2000) Conceptualizing and measuring peer acceptance and rejection, in A.H.N. Cillessen (Ed.), *New Directions for Child and Adolescent Development* (Vol. 88), pp. 3–10. San Francisco, CA: Jossey-Bass.

Cliff, N. (1996) Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, **31**, 3, pp. 331–350.

Coie, J.D., Dodge, K.A. & Coppotelli, H. (1982) Dimensions and types of social status: A cross-age perspective. *Developmental Psychology*, **18**, 4, pp. 557–570.

Dunn, J.C., Dunn, J.G.H. & Bayduza, A. (2007) Perceived athletic competence, sociometric status, and loneliness in elementary school children. *Journal of Sport Behavior*, **30**, 3, pp. 249–269.

Durkin, M. (2007) On e-banking adoption: From banker perception to customer reality. *Journal of Strategic Marketing*, **15**, 2, pp. 237–252.

Edmonds, J. (2008) *How to Think About Algorithms*. New York, NY: Cambridge University Press.

Feather, N.T. (1973) The measurement of values: Effects of different assessment procedures. *Australian Journal of Psychology*, **25**, 3, pp. 221–231.

Finn, A. & Louviere, J.J. (1992) Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy and Marketing*, **11**, 1, pp. 12–25.

Gagne, F. (1998) The prevalence of gifted, talented, and multitalented individuals: Estimates from peer and teacher nominations, in R.C. Friedman & K.B. Rogers (Eds.), *Talent in Context: Historical and Social Perspectives on Giftedness*, Washington, D.C.: American Psychological Association

Henry, D.B. (2006) Associations between peer nominations, teacher ratings, self-reports, and observations of malicious and disruptive behavior. *Assessment*, **13**, 3, pp. 241–252.

Ipeirotis, P. (2008) *Mechanical Turk: The Demographics*. Retrieved 15 February 2010 from http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html

Kamakura, W.A., & Mazzon, J.A. (1991) Value segmentation: A model for the measurement of values and value systems. *Journal of Consumer Research*, **18**, 2, pp. 208–218.
Klein, M., Dülmer, H., Ohr, D., Quandt, M. & Rosar, U. (2004) Response sets in the measurement of values: A comparison of rating and ranking procedures. *International Journal of Public Opinion Research*, **16**, 4, pp. 474–484.
Krosnick, J.A. & Alwin, D.F. (1988) A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, **52**, 4, pp. 526–538.
Louviere, J.J. (1991) *Best-worst Scaling: A Model for the Largest Difference Judgments*, Working Paper, University of Alberta.
McCarty, J.A. & Shrum, L.J. (1997) Measuring the importance of positive constructs: A test of alternative rating procedures. *Marketing Letters*, **8**, 2, pp. 239–250.
Michalowicz, J.V., Nichols, J.M. & Bucholtz, F. (2013) *Handbook of Differential Entropy.* Boca Raton, FL: CRC Press.
Munson, J.M. & McIntyre, S.H. (1979) Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research*, **16**, 1, pp. 48–52.
Newcomb, A.F. & Bukowski, W.M. (1983) Social impact and social preference as determinants of children's peer group status. *Developmental Psychology*, **19**, 6, pp. 856–867.
Ovadia, S. (2004) Ratings and rankings: reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology*, **7**, 5, pp. 403–414.
Peery, J.C. (1979) Popular, amiable, isolated, rejected: A reconceptualization of sociometric status is preschool children. *Child Development*, **50**, 4, pp. 1231–1234.
Sawtooth Software (2013) *The MaxDiff System Technical Paper*. Retrieved 2 June 2015 from http://www.sawtoothsoftware.com/support/technical-papers/maxdiff-best-worst-scaling/maxdiff-technical-paper-2013
Schwarzwald, J., Koslowsky, M. & Mager-Bibi, T. (1999) Peer ratings versus peer nominations during training as predictors of actual performance criteria. *The Journal of Applied Behavioral Science*, **35**, 3, pp. 360–372.
Shannon, C.E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423, 623–656.
Simon, V.A., Aikins, J.W. & Prinstein, M.J. (2008) Romantic partner selection and socialization during early adolescence. *Child Development*, **79**, 6, pp. 1676–1692.
Vanleeuwen, D.M. & Mandabach, K.H. (2002) A note on the reliability of ranked items. *Sociological Methods & Research*, **31**, 1, pp. 87–105.
Wilkie, D. (1980) Pictorial representation of Kendall's rank correlation coefficient. *Teaching Statistics*, **2**, 3, pp. 76–78.

Table 1. HL3, HL5, and Full Ranking generate highly correlated results.

Actual	hl3	hl5	Full Ranking
California	Texas	Texas	California
Texas	California	Florida	Texas
New York	New York	California	New York
Florida	Florida	New York	Florida
Illinois	Illinois	Illinois	Illinois
Pennsylvania	Pennsylvania	Pennsylvania	Georgia
Ohio	New Jersey	New Jersey	Pennsylvania
Michigan	Massachusetts	Ohio	North Carolina
New Jersey	Minnesota	Georgia	Massachusetts
Georgia	Michigan	North Carolina	South Carolina
North Carolina	Georgia	Virginia	Arizona
Virginia	Virginia	Washington	Ohio
Massachusetts	South Carolina	Minnesota	Washington
Indiana	Ohio	South Carolina	Tennessee
Washington	Arizona	Michigan	Michigan
Tennessee	Washington	Louisiana	Virginia
Missouri	Alabama	Tennessee	New Jersey
Wisconsin	Colorado	Massachusetts	Nevada
Maryland	Mississippi	Colorado	Louisiana
Arizona	North Carolina	Mississippi	Minnesota
Minnesota	Indiana	Indiana	Maryland
Louisiana	Tennessee	Arizona	Indiana
Alabama	Louisiana	New Mexico	Alabama
Colorado	Kansas	Wisconsin	Missouri
Kentucky	Missouri	Kentucky	Wisconsin
South Carolina	Maryland	Alabama	West Virginia
Oklahoma	Oklahoma	Missouri	Connecticut
Oregon	Arkansas	Connecticut	Oregon
Connecticut	New Mexico	Maryland	Mississippi
Iowa	West Virginia	Nebraska	Colorado
Mississippi	Kentucky	Kansas	Utah
Kansas	Oregon	Arkansas	New Mexico
Arkansas	Maine	Oklahoma	Kentucky
Utah	Wisconsin	West Virginia	New Hampshire
Nevada	Connecticut	Nevada	Arkansas
New Mexico	Nevada	Iowa	Kansas
West Virginia	Nebraska	Idaho	Oklahoma
Nebraska	Utah	Maine	Nebraska
Idaho	Iowa	Oregon	Maine

Maine	New Hampshire	Hawaii	Idaho
New Hampshire	Delaware	New Hampshire	Iowa
Hawaii	Vermont	Delaware	Delaware
Rhode Island	Idaho	Utah	Hawaii
Montana	Wyoming	Vermont	Vermont
Delaware	Rhode Island	Rhode Island	Montana
South Dakota	Montana	Wyoming	North Dakota
North Dakota	South Dakota	Montana	Rhode Island
Alaska	Hawaii	Alaska	Alaska
Vermont	North Dakota	South Dakota	Wyoming
Wyoming	Alaska	North Dakota	South Dakota

Table 2. High-Low partial rankings correlate with Max-Diff results and are easier for respondents to complete.

Fruit	MaxDiff Utilities	MaxDiff Rank	Most Favorite	Least Favorite	HL Net	HL Net Rank
Strawberry	8.45	1	67	6	61	1
Pineapple	6.81	2	44	11	33	3
Cherry	6.74	3	16	12	4	12
Peach	6.56	4	34	9	25	7
Raspberry	5.89	5	30	16	14	9
Watermelon	5.86	6	46	12	34	2
Banana	5.16	7	45	12	33	3
Grape	5.14	8	23	9	14	9
Orange	5.00	9	34	7	27	5
Blueberry	4.68	10	32	10	22	8
Apple	4.53	11	36	9	27	5
Tangerine	3.76	12	9	16	-7	14
Cantaloupe	3.75	13	16	30	-14	16
Pear	3.70	14	16	16	0	13
Plum	3.03	15	8	24	-16	17
Mango	2.98	16	28	16	12	11
Kiwi	2.88	17	8	27	-19	18
Honeydew	2.71	18	11	39	-28	20
Apricot	2.40	19	4	37	-33	23
Lemon	2.27	20	8	18	-10	15
Avocado	1.84	21	17	47	-30	21
Grapefruit	1.82	22	9	39	-30	21
Lime	1.80	23	6	27	-21	19
Papaya	1.15	24	2	44	-42	24
Date	1.06	25	7	63	-56	25

Figure 1. The time to complete the High-Low task is nearly independent of length list.

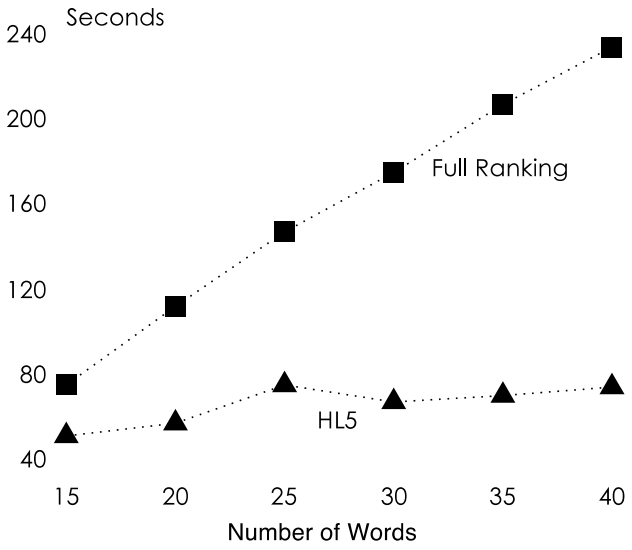

Figure 2. HL5 accuracy decreases on long lists.

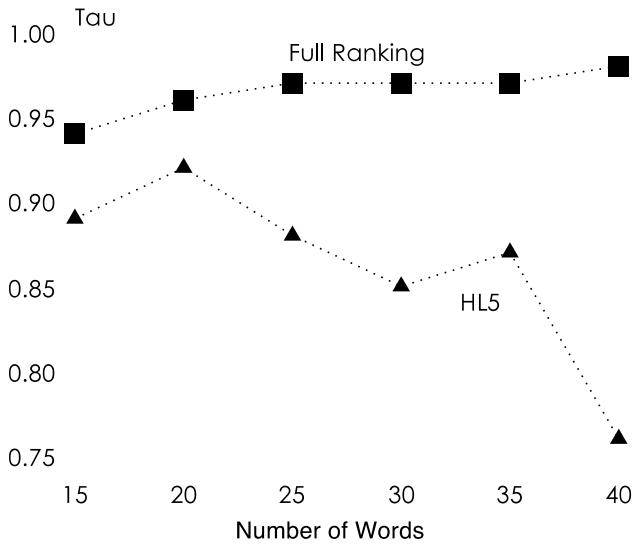

Figure 3. HL methods provide same accuracy as full ranking but with less effort.