

# Quality Control in Crowdsourcing: An Objective Measurement Approach to Identifying and Correcting Rater Effects in the Social Evaluation of Products and Services

## Rense Lange

Integrated Knowledge Systems  
107 Carefree Drive  
Chatham, Illinois 62629  
renselange@gmail.com

## Xavier Lange

Tureus Inc  
530 Genter St  
La Jolla, California 92037  
xrlange@tureus.com

### Abstract

Crowdsourcing requires new strategies to evaluate the workers involved as well as the quality of workers' output. Using customer feedback data, we introduce multi-faceted Rasch scaling as an evaluation technique to assess the contributions of workers and products simultaneously within a single coherent measurement framework. Based on a data set of about 250,000 customers who rated nearly 115,000 products, for a total of nearly 3 million cases, we found that product ratings reflect almost as much the existence of stable rater differences as they are indicative of the products' properties. We illustrate how Rasch scaling provides extensive quality control mechanisms; as well we show how aberrant workers and products can be identified so that appropriate feedback and/or corrective actions can be initiated.

### Introduction

The online gathering of customer feedback is arguably one of the earliest forms of crowdsourcing, as soliciting and

displaying reviews helps buyers and sellers alike in their decision-making. The growth of services like Amazon's Mechanical Turk and CrowdFlower indicates that crowdsourcing can be used for increasingly diverse efforts in the large-scale gathering and integration of human judgments. Yet, the greatest benefit of this approach – i.e., its flexibility and the possibility of having large numbers of workers participate online – is also its greatest potential weakness, precisely because the traditional worker-supervisor interactions are absent from crowdsourcing. In addition to requiring new approaches for recruiting, selecting, training, and rewarding potential workers, crowdsourcing also needs new strategies for evaluating its results. This is further complicated by the fact that crowdsourced tasks by definition have no obvious “right” answer, or at least not an answer that can be computed or otherwise obtained without invoking human judgments. These considerations apply to customer evaluations and crowdsourcing alike, and lessons learned in one area thus inform the other as well.

This paper takes customer evaluations as a model because this is one of the most widely used forms of crowd sourcing, and large databases are available with reviewers' ratings. The data analyzed here consist of the online reviews by about 250,000 customers who voluntarily rated

nearly 115,000 products available at the Amazon.com website, by awarding one through five “stars” (five stars being best). Rather than evaluating workers / raters and products in isolation, a probabilistic “objective measurement” approach in the form of Rasch (1960/1980) scaling is used. Rasch scaling simultaneously scales the contributions of workers and products on a common dimension, and it has excellent quality control mechanisms. Rasch scaling has long been used in education, psychology, and medicine (an overview can be found in Bond and Fox, 2007). The approach has seen little use in artificial intelligence, but it was successfully adopted by Lange, Greiff, Moran, and Ferro (2004) to evaluate the linguistic AI systems across different tasks in the Question Answering track of the Text REtrieval Competition (TREC). This paper expands that research by including a mediating “facet,” the human raters.

### Multi-facetted Rasch Scaling

It is customary to treat ordinal information, including product ratings or “stars”, as if these represented true quantitative values for which it is justified to compute statistical indices like means, standard deviations, etc. Yet, it has long been noted (see e.g., Churchill, 1979; Wright and Masters, 1982) that there are good reasons to treat such information with due caution. For instance, rating scales provide ordinal information at best and translating ordinal information into quantitative measures requires additional steps and assumptions (Wright and Masters, 1982). Second, even highly qualified and trained professional evaluators cannot make ordinal ratings consistently. For instance, educational assessment companies routinely train raters (often experienced teachers) to evaluate thousands of student essays written for statewide educational tests. Using overlapping tasks, the performance of these raters is compared to that of their colleagues to establish reliability and to model papers graded by expert raters to establish validity. Yet, despite the considerable training, feedback, and possibly, retraining, the educational raters continue to differ with respect to their severity / leniency, and exact agreement rates of 65% are the norm in this industry.

Customers are not trained in the use of product ratings, their participation depends on diverse motivations, and thus disagreements are only to be expected. Moreover, customers’ idiosyncrasies will go unchecked, and any leniency / severity differences are probably stronger among such volunteers than among rating professionals. To assess the impact of stable rater differences and product quality simultaneously, we use multi-facetted Rasch scaling (Linacre, 1989). In this approach, the number of stars awarded by the Amazon customers is modeled probabilistically as the outcome of the interaction between

three elements: the perceived quality of the product ( $B_n$ ), the leniency/ severity of the reviewer ( $C_j$ ), and the boundaries between the numbers of “stars” being awarded,  $F_k$ . A compact way to present this model is in terms of the log-odds defined by the probability of observing  $k$  stars being awarded ( $P_{njik}$ ) vs. that of awarding  $k-1$  stars ( $P_{(n-1)jk}$ ):

$$\ln \left( \frac{P_{njik}}{P_{(n-1)jk}} \right) = B_n - C_j - F_k, \quad (1)$$

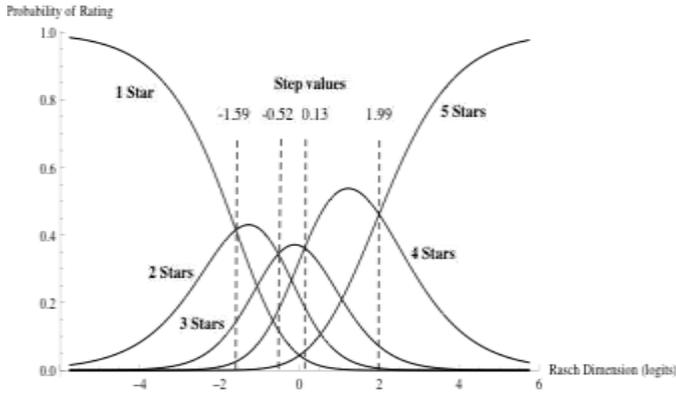
In Eq. 1 the “step” values  $F_k$  define the boundary points at which observing  $k$  stars has the same likelihood as observing  $k-1$  stars. Note that B, C, and F are expressed in a common metric, the log-odds, or logit, scale as is defined by the ratio on the left.

Rasch measurement has the property that all model parameters have sufficient statistics; moreover, these statistics are *always* one-to-one (but non-linear) functions of the observable raw data summed over the appropriate (sub-) facets. Thus, as will become relevant later, rank-correlation coefficients rather than standard (i.e., linear) Pearson correlation coefficients should be used to evaluate the “raw-to-Rasch” translations. The raw-score to parameter translations can be estimated using Linacre’s (2011) Facets software. This software also produces model-fit information for each rater and product in the form of an “Outfit” statistic, which is a normalized  $\chi^2$  statistic with an expected value of 1. Excursions below 1.0 are typically unproblematic, and values in the range 1.0 to 1.6 – while indicative of noise – can be tolerated. Outfit values greater than 2.0 indicate that measurement is distorted (Bond and Fox, 2007). Misfitting products and/or raters can be omitted to avoid measurement issues. However, rater and product misfit can also be seen as a learning opportunity to rethink the task and/or to initiate rater training.

Although the present research uses a large data set, we note that Rasch scaling is also quite suitable for small scale applications with no more than a few hundred cases (see, Bond and Fox, 2007). For instance, medical applications can be cited (e.g., Pesudovs, Garamendi, Keeves, and Elliot, 2003) where meaningful results were obtained based on the data of just forty-three patients who were rated on 22 features. However, rather than focusing on sample sizes, we recommend that data is gathered to reach acceptable (and application specific) accuracy. The Facets software assists in this respect as it computes the standard error of all model parameters.

Obtaining meaningful estimates requires that the parameters in Eq. 1 are uniquely identifiable. Engelhard (1997) showed that a sufficient condition for uniqueness is that the response strings for every pair of customers (P, Q) and every pair of products (R, S) contain at least one

response where P (R) is higher than Q (S) and at least one response where Q (S) is higher than P (R). Such relations must either hold directly or indirectly through the response strings of their pairings with other raters or products. The Facets software identifies all maximal subsets of raters and products that are sufficiently connected to guarantee uniqueness. We expected that multiple subsets will occur in the present data as many consumers write just a single product review.



**Figure 1:** Modeled category structure of evaluations in terms of the probability (Y-axis) of observing 1 through 5 “stars” being awarded, given the Rasch dimension (X-axis).

### Preliminaries

A database of 5,148,669 cases was available containing the judgments by 1,128,997 reviewers concerning 149,899 different Amazon products. As was anticipated, several disjoint subsets of raters and products were discovered. Yet, an extremely large subset of 2,951,061 cases was identified (i.e., 57% of all cases) for which the product, rater, and rating step parameters could be unambiguously estimated. This subset covers 113,910 different products as rated by 268,381 different reviewers. For space reasons, the following does not distinguish between particular subcategories of products or reviewers. The Facets 3.68.1 software (Linacre, 2011) was used throughout. The main run took approximately 10.5hrs on a Mac Pro with 2 x 2.8 GHz Quad-Core Intel Xeon under Max OS X V. 10.7.1, which used VM Fusion to emulate 32-bit Windows 7 Professional with 4.76 GB of RAM.

## Results

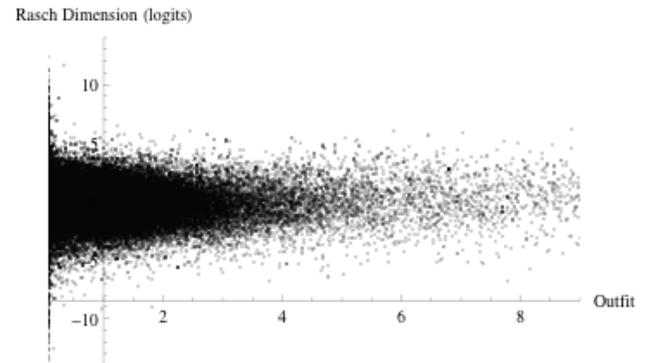
Figure 1 plots the modeled probabilities of awarding either 1, 2, ..., or 5 stars along Y-axis, as a function of the magnitude of the Rasch variable (X-axis) defined by the difference between rater and product measures (see Eq. 1). As is the case in nearly all applications, the step transitions

are spaced quite unevenly. For instance, the distance between the second and third steps ( $0.13 + 0.52 = 0.65$  logits) is less than half of that between steps 3 and 4 (1.76 logits). Thus, directly contradicting the notion that raw scores represent quantitative data, awarding three rather than two stars requires a far smaller increase in perceived product quality than does the change from four to five stars.

Figure 2 shows the parameter values of the nearly 114,000 products along the Y-axis together with their Outfit (X-axis). It can be seen that the product evaluations ( $M_B = 0.0, SD_B = 1.95$ ) cover a wide range of values, as some products are perceived as highly positive and others as highly negative. There is no evidence of bimodality and the product distribution does not deviate significantly from a normal curve ( $p > 0.10$ ). Also, the high value of the reliability index (0.92) indicates that products’ evaluations were estimated consistently. In this context, reliability is defined as follows:

$$\frac{\text{Measure variance}_{\text{observed}} - \text{Error variance}_{\text{Rasch}}}{\text{Measure variance}_{\text{observed}}} \quad (2)$$

Note that the the large right (and truncated) tail of high Outfit values in Figure 2 indicates that reviewers did not always agree. The Outfit statistics of about 12% of the products exceeded 2.0, indicating that their measurements are distorted – i.e., some high (low) scoring products received a disproportionate number of low (high) ratings. Another 6% had Outfits in the range 1.6 to 2.0 and these products’ evaluation parameters should not be accepted at face value.



**Figure 2:** Products’ evaluations along Rasch dimension (Y) by their Outfit values (X)

We further note that the Outfit of nearly 10% of the raters exceeded 2.0, indicating that these raters showed unacceptable noise in their judgments, and another 4% showed an Outfit in the range 1.6 to 2.0. These findings reflect that some customers’ ratings were too noisy (i.e.,

too high or too low) given their severity / leniency parameters and/or the ratings products received from others. Accordingly, a picture very similar to Figure 2 emerges when customers' severity / leniency parameters are plotted against their Outfit values (to conserve space this figure is not included). A crucial finding is that customers' severity / leniency shows considerable variation ( $M_C = -1.27$ ,  $SD_C = 1.73$ ) indicating that this factor is almost as powerful as the products themselves in determining consumers' ratings.

Another way to quantify the impact of raters' severity / leniency is to note that in the absence of rater differences, Equation 1 implies that the average number of stars awarded should show a perfect rank correlation (i.e.,  $r = 1$ ) with products' estimated  $B_j$ . This is clearly not the case here as the Spearman rho rank correlation is only 0.78, indicating that respondents' scores explained no more than 60% of the variance in their product evaluations. We already noted that raters' evaluation data are ordinal at best since the spacing of the category boundaries is unequal. To this can now be added that ignoring the rater severity ( $C_j$  in Equation 1) leads to a significant deterioration in these ordinal properties to the point that one can no longer be sure that a judgment of, say, '\*\*\*\*' reflects a more positive product evaluation than does '\*\*\*'.

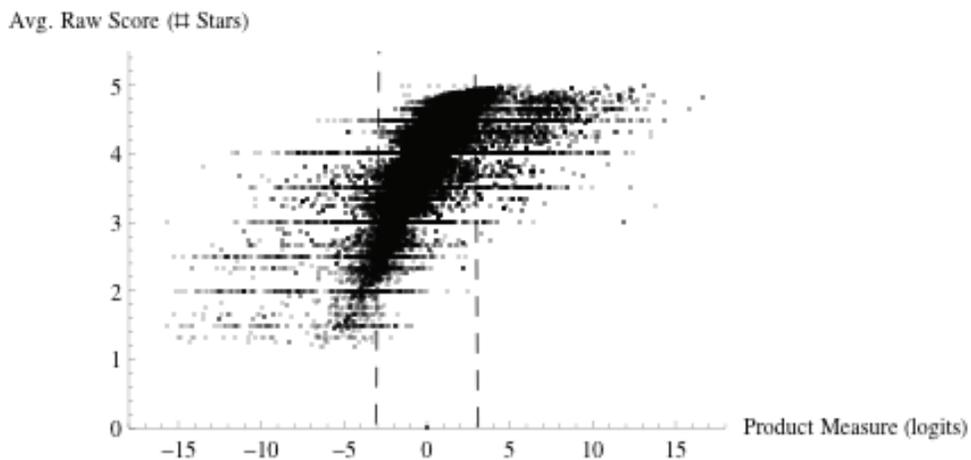
An important consequence of these findings is that the use of direct agreement between raters' judgments may be a poor or misleading index of rater consistency or reliability. Instead, we propose that using the dispersion (variance) of the distribution of rater severity / leniency parameters is to be preferred for quality control purposes.

It can be seen that the relation between these two variables is decidedly non-linear because raw differences near the top are underestimate the true differences, thereby again calling into question the use of raw scores (presently, the number of stars) as quantitative indices. Also, consider a hypothetical product with a perceived quality of 0.0 logits, and imagine that this product being rated either by severe raters (leftmost dotted vertical line at the 5<sup>th</sup> percentile in the rater distribution = -2.95 logits) or by lenient raters (rightmost vertical dotted line at the 95<sup>th</sup> percentile = 2.95 logits). Then, severe customer's ratings' would cover a range from about 1.5 to 3.6 "stars" along the Y-axis, while lenient customers might cover the range from about 3.5 to 4.9 stars.

The reliability of the rater estimates is 0.83, indicating that the observed rater differences reflect stable individual characteristics rather than chance variation. Thus, while raters are quite consistent across occasions and products are relatively stable across raters, the considerable variation in the products' ratings is to a significant extent the product of consumers' severity / leniency rather than of the perceived quality of the goods and services they rated.

## Summary and Discussion

We have outlined and illustrated the use of multi-faceted Rasch scaling in crowdsourcing, as embodied in consumers' online rating of Amazon merchandise. The chosen example was realistic in that large numbers of consumers and products were scaled simultaneously within a single coherent framework, and we've demonstrated that



**Figure 3:** Raw scores (Y-axis) vs. Logits (X-axis)

The implications of this finding are illustrated in Figure 3, which plots the average number of stars awarded (Y-axis) as function of the Rasch product parameters (X-axis).

processing over 5 million cases is feasible. Future research can be finer grained because no distinctions were made among customer types and/or the nature of the products

being reviewed. When required, Linacre's (2011) Facets software can accommodate additional independent variables like customer demographics and/or product categories by including these factors into Eq. 1 (Linacre, 1989). Interactions can be modeled as well; we might ask, for instance, whether women differ from men depending on the nature of the products being rated or the rating dimensions being used.

It should be noted that multi-faceted Rasch scaling can also be used with much smaller data samples than we had available here. As well, the approach is not limited to using a single rating dimension, as multiple response dimensions can be used simultaneously, or the same object can be rated repeatedly on multiple occasions. At the same time, it is possible to include additional independent variables as needed. For instance, one might include date information to study product evaluation trends over time, or rater demographics to aid in product targeting. Finally, when needed, the current approach applies to situations where workers evaluate each other (including a "360"-style), or where workers' job is to produce paired comparisons (i.e., judgments of the form  $A > B$ , vs.  $B > A$ ), including paired comparisons with graded categories (e.g.,  $A \gg B$ ,  $A > B$ ,  $A = B$ ,  $B > A$ ,  $B \gg A$ ).

Our most important finding is that product ratings reflect almost as much the existence of stable rater differences as they are indicative of the products' perceived quality. As random assignment of raters is rarely feasible, this means that crowdsourcing can produce high quality results only if raters' severity/ leniency is explicitly measured and compensated for. Rasch scaling has the additional advantage that misfitting elements can be identified, and thus outliers can be deleted so as to protect the integrity of the crowdsourcing effort. Moreover, it is often possible to exploit products' and raters' misfit dynamically as a learning opportunity to initiate corrective actions. For instance, it may be that some raters misunderstood the task or they might simply be malingering, some products may be unfamiliar to the raters thereby yielding seemingly random responses, or the rating dimension might be ambiguous. In such cases early intervention would serve to limit unnecessary expense.

We wish to stress that the multi-faceted Rasch approach is *not* limited to subjective tasks like product evaluation, or other tasks that rely heavily on social consensus. For instance, during the crowdsourcing of the line-labeling of images (Sorokin and Forsyth, 2008) one could present workers with images that had been expertly labeled previously with great precision, and new labels could be scored in terms of a suitably defined "distance" to such ground-truth stimuli. We noted earlier that the estimation process requires that all parameters be uniquely identifiable and that thus some tasks have to be presented repeatedly. Fortunately, minimal overlap often suffices and

complicated rotation and assignment plans are not required. For instance, Linacre (1997) describe a randomized plan which achieves full parameter identifiability while using just 8% of all possible combinations of tasks and raters.

## Future Work

To decrease detection latency and to facilitate intervention, research is currently underway to optimize and automate multi-faceted Rasch analyses of crowdsourcing data by updating the model parameters' estimates and fit indices in real time. When dealing with relatively small data sets, Facets can simply be re-run over all available data as needed. However, this approach becomes cumbersome in applications when large quantities of new data arrive continuously and where it is of utmost importance to identify aberrant raters and products/tasks as soon as possible. For this reason we are currently implementing a Rasch scaling system that analyzes the data it receives dynamically.

We have found that the "Unconditional Likelihood" (UCON) parameter estimation approach (see, e.g., Wright and Masters, 1982; Linacre, 1989) already used by the Facets software can suitably be adapted to this purpose. However, important statistical issues arise in practice. For instance, when fit to the model is monitored continuously, it is only to be expected that some raters and tasks will (erroneously) appear to show misfit by chance alone. Thus, dynamic versions of the misfit criteria listed in earlier sections will have to be formulated. Also, rather than pairing raters and tasks essentially at random or based on availability, there is much to be gained by combining raters and tasks in ways that maximize the information contents of the raters' ratings.

Identifying optimal facet combinations can be modeled as a generalization of computer adaptive testing (CAT) techniques, which are increasingly used in psychological and educational testing (Lange, 2007). Traditional CAT attempts to maximize the information gained from the next observation by identifying optimal combinations of persons and tasks. In the present context, this approach is complicated by the fact that more than two facets are involved. At the same time, the Rasch model parameters are not known reliably, but they are refined dynamically, while the underlying model is not (yet) identifiable. To satisfy all the different – and possibly competing – demands simultaneously we suggest using linear programming methods in CAT along the lines pioneered by van der Linden (2000).

Finally, the system should provide a variety of tools to foster quality control in crowdsourcing. The single most important step in quality control probably is worker

selection, training, and monitoring, but these topics fall outside the scope of the present paper. However, given that workers have been employed and are performing a judgment task, the system should monitor the following threats to reliability and validity:

- **Rater severity / leniency.** As we have demonstrated, workers differ systematically with respect to the overall severity or leniency in expressing their judgments. Rasch scaling solves this problem by explicitly taking this factor into account. Experience indicates that such severity / leniency differences persist despite feedback and retraining efforts. Moreover, attempts to remove, correct, or minimize leniency differences may be counter-productive because workers' performance may become unstable as a result (e.g., due to inability/ unwillingness to comply, disagreement, or fear of being fired).
- **Bias.** Even when faced with the same stimulus, workers' judgments on specific tasks may vary systematically with the demographic subgroups to which they belong. Such biases can occur at the individual or at the group level. Note that biases operate in addition to differences in worker severity/leniency (if any). Using statistical terminology, biases are thought of as interactions rather than as main effects.
- **Idiosyncrasies.** Other threats to reliability/validity include: A tendency to use either extreme or central judgments (*i.e.*, *black-white vs. non-committal judgments*), "Halo" effects or stereotypes (*a carryover of one evaluation into others: I believe that all X are Y, therefore this X must be Y also*), and "Playing it safe" (*e.g.*, *the over-use of middle categories to minimize inter-rater disagreement*), and Inattention / instability.

## References

Bond, T.G., and Fox, C.M. 2007. *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Churchill, G.A. 1979. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16, 1, 64-73.

Engelhard, G. 1997. Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19-33.

Lange, R. 2007. Binary Items and Beyond: A Simulation of Computer Adaptive Testing Using the Rasch Partial Credit Model. In: Smith, E. and Smith, R. (Eds.) *Rasch Measurement: Advanced and Specialized Applications*. Pp. 148-180, Maple Grove, MN: JAM Press.

Lange, R., Moran, J., Greiff, W.R., and Ferro, L. 2004. A Probabilistic Rasch Analysis of Question Answering Evaluations. *HLT-NAACL 2004*: 65-72.

Linacre, J. M. 1989. *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J.M. 1997. *Judging plans and facets*. Mesa Research Note #3. (consulted on September 4, 2011, <http://www.rasch.org/rn3.htm>)

Linacre, J. M. 2011 *Facets computer program for many-facet Rasch measurement, version 3.68.1*. Beaverton, Oregon: Winsteps.com.

Pesudovs, K., Garamendi, E., Keeves, J.P., and Elliot, D.B. 2003. ADVS for Cataract Surgery Outcomes. *Investigative Ophthalmology & Visual Science*, 44, 2892-2899.

Rasch, G. 1960/1980. *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Sorokin, A., and Forsyth, D. 2008: Utility data annotation with amazon mechanical Turk. *First IEEE Workshop on Internet Vision at CVPR 08, Anchorage, Alaska*.

van der Linden, W. J. 2000. Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston: Kluwer.

Wright, B. D., & Masters, G. N. 1982. *Rating scale analysis*. Chicago: MESA Press.