Rating a Subset of Items in Comparison to a Superset

Abstract

What happens when people are ask to rate only a subset of items? Does rating only a subset of items produce biased values compared with rating all the items? Nine hundred and seventy participants were asked to provide comparison ratings for a subset (1, 2, 3, 4, 5, or all 10) of pieces of art from a set of ten pieces. Participants' ratings of these subsets were similar to when rating all items. There was a small negative bias when rating only one piece, but overall mean ratings for all conditions were near the scales' mid-points. The results are encouraging. They suggest that researchers may be able to have participants rate only a subset of items, instead of having to rate all the items, without negatively affecting results (with some caution if participants are rating only one item).

Keywords: estimating effectiveness, judgement, ratings, subset

Rating a Subset of Items in Comparison to a Superset

What happens when someone is asked to make ratings about a small subset of items in comparison with the larger superset from which the subset is drawn? Do they tend to rate this subset higher or lower than if they had rated all the items? This is an important applied question because academic and marketing researchers often want to know how people feel about a large number of items, but asking people about all of these items would be impractical.

There are several applications where people are only asked about a subset of items. For example, in market research often a company wants to know respondents' opinions about dozens of products, but knows that asking about each would greatly lengthen the survey and cause respondent fatigue. The literature is mixed about how long a survey can be before substantial fatigue negatively affects results (e.g., Bradley & Daly, 1994; Hess et al., 2012). For some psychology studies the numbers of stimuli that are of interest can be substantial. For example, if wanting to present several scenarios which systematically vary on say five dimensions each with four possible values, the total number of scenarios is $5^4 = 625$. In these cases fractional factorial designs (see Box et al., 2005, Chapters 6–7) are used so that each respondent is presented with a reasonable number of scenarios to judge and together the analyst can still estimate main effects and interactions that were deemed important, but the estimation of other interactions is confounded.

Two rating procedures are used in the current study. These are: having participants provide ranks $1 \ldots n$ for the item and having them rate the stimuli on a 0–100 scale. There have been several arguments in preference for each of these approaches. Some are theoretical based on the cognitive tasks required (e.g., Klein et al., 2004) and some based on large scale surveys (e.g., Alwin & Krosnick, 1985). Alwin & Krosnick compared these procedures using the General Social Survey and found the two produced the same general orderings of preferences. Russell & Gray (1994) compared rankings and ratings using pieces of art. They found the two produced similar results. Here these two procedures are used to

test for the generality of the main results. Following Russell & Gray, pieces of abstract art are used in the current study.

The study was conducted online using the Qualtrics software (`qualtrics.com`). The sample was recruited and compensated using Amazon's Mechanical Turk (MTurk). Several studies have compared the results of studies using MTurk and laboratory methods (Crump et al., 2013). While some differences occur, an oft-cited conclusion is that "the data obtained are at least as reliable as those obtained via traditional methods" (Buhrmester et al., 2011, p. 3). The design of the current study followed guidelines for effective MTurk use by behavioral scientists (e.g., Mason & Suri, 2012; Sheehan & Pittman, 2016). MTurk is particularly well-suited when needing a large sample of people to do a brief task.

Research Questions

There are two main research questions (RQs). When evaluating these both the unadjusted and the Holm adjusted p-values (for these two) will be reported.

RQ 1 Is there a main effect difference depending on whether participants rate a subset (size: 1, 2, 3, 4, 5) of items or all ten of the items?

RQ 2 Is there a difference between rating just one item versus rating a subset of multiple items?

There are several additional ancillary research questions. While these are not central to the focus of the study, they are valuable to report.

RQ 3 Overall, is the mean response at the mid-points for the scales?

RQ 4 Is there a main effect of scale, with one of the approaches (ranking and rating) producing higher or lower estimates (after the responses have been put on the same scale)?

RQ 5 Does the accuracy of the ratings vary by whether 2, 3, 4, or 5 items are rated?

The final set of hypotheses concern interactions and the generality of any results found above.

RQ 6 Is there an interaction between scale and the subset versus rating all conditions?

RQ 7 Is there an interaction between scale and whether one or several items of a subset are rated?

RQ 8 Does the scale moderate any of the effects found in #RQ 5?

RQ 9 Are there demographics differences ...

a. by age (main effects and interactions)?

b. by education level (main effects and interactions)?

c. by gender (main effects and interactions)?

Methods

Sample

The sample size was chosen based on the power to detect two effects (RQ 1 and RQ 2) if they are small in Cohen's (1992) terms ($.2\sigma$). The critical α level for the power analysis was set to .025 because a conservative approach (Bonferroni's method) to adjusting for multiple p-values is dividing by the number of tests, here two. In the results section Holm's method will be used since it is maintains familywise Type I error, but is more powerful than Bonferroni's method. Different statistical procedures will be used, but to give an approximate suggestion for an appropriate sample size G*Power (Faul et al., 2007) for comparing two independent means estimated that a sample of 954 would achieve 80% power (the R command `power.t.test(delta = .2,sig.level=.025,power=.8)` can also be used). Given the likelihood that some data would need to be excluded one thousand participants were recruited.

The study was published on Amazon Mechanical Turk (AMT) on Tuesday, November 27, 2018. Following the guidelines from Sheehan & Pittman (2016), only people 18 years of age and over in the US who had successfully completed 500 AMT tasks with a 97% acceptance rate[1] were allowed to sign up. Some responses came from duplicate IP

[1]AMT allows requesters to review each participant's responses to decide whether to pay them (in AMT terminology to "accept" them). If a participant does not spend ample effort on these tasks it is likely the

addresses. While the second use of an IP address would be from a different Amazon account (and according to AMT rules from a different person), the people may have talked about the task and therefore the second person was excluded. Some (14, or 1.42%) of the remaining participants responded faster than 20 seconds and were excluded, though still paid. The final sample size is 970.

Characteristics of typical AMT samples are discussed in many sources. Sheehan & Pittman (2016, Ch. 2) summarize much of this research: AMT samples tend to be younger than the general population, have a range of education levels, and approximately half are female. They tend to be more diverse than the typical psychology laboratory sample. Three demographic variables (age, education level, and gender) were asked at the end of this study. Participants in this sample have similar demographics to those reported by Sheehan & Pittman (2016, p. 17). Participants were asked for their year of birth. Binning these into decades: 1.45% were born in the 1940s, 5.79% in the 1950s, 13.03% in the 1960s, 19.65% in the 1970s, 39.61% in the 1980s, and 20.48% in the 1990s. Participants reported their education level: 12.99% had an associate degree in college (2-year), 38.14% had a bachelor's degree in college (4-year), 1.55% had a doctoral degree, 11.86% had a high school graduate (high school diploma or equivalent including GED), 0.31% had less than high school degree, 9.18% had a master's degree, 1.03% had a professional degree (JD, MD), and 24.85% had some college but no degree. Participants' self-reported genders were: 50.21% female, 49.38% male, 0.10% other, and 0.21% ticked "prefer not to specify."

Stimuli

There is a large amount of social psychology research about how person ratings can be affected by non-diagnostics aspects of the person descriptions (e.g., Fiske & Neuberg, 1990). Instead of rating people, following Russell & Gray (1994), participants rated pieces of abstract art. Ten pieces, all created by the same artist (Soni Wright,

---

person would have a low acceptance rate so could not have taken part in the current study.

www.ojaistudioartists.org/soni-wright), were chosen and are shown in Figure 1. None of these pieces were displayed in juried shows outside of Southern California and none of the participants are likely to have seen any of them.

Procedure

Participants were recruited on AMT and if they agreed they went to the Qualtrics survey (housed on a Qualtrics server). The first screen thanked them for volunteering and displayed the ten pieces of art. The next screen showed participants these ten again and one of them (chosen at random) to compare with the whole set. A random half (50.21%) were asked to provide an integer from 1–10 corresponding to whether the piece of art was the best (#1), the worst (#10), and so on. The system required that participants entered an integer before allowing them to progress. Half (49.79%) were asked to do this on a 0–100 rating scale with numeric labels at each ten-points, the words "Worst" and "Best" at the two extremes, and "Average" at the midpoint. Participants moved a bar, which started at the midpoint, to the their desired location. The system required that the participant moved the bar before progressing.Participants were randomly allocated to rate 1, 2, 3, 4, 5, or 10 pieces of art, with three times as many sampled to see just one piece, and twice as many to see just two pieces. This over-sampling was because participants rating only one or two pieces of art provide less information than those rating more pieces of art. Thus, this is a 2×6 unbalanced between subject design. The choice of which art pieces to display and the order in which to display them were based on Qualtrics in-built randomizer.

After completing the ratings participants answered the three demographic questions (year of birth, highest education level, and gender). They were thanked and given a code to use on the AMT webpage to get paid. Payment was \$0.50. This was based on a conservative estimate for the amount of time (three minutes) to complete the task and Sheehan & Pittman's (2016) guidelines of paying \$0.15 per minute. Ninety six percent of those rating ten items finished in less than three minutes. All participants were paid. This

study was approved by the Alder Institutional Review Board.

Results

Descriptive Statistics

To compare responses on the two scales, responses on each were transformed so that -1 corresponds to the lowest possible rating, 0 to the mid-point, and +1 to the highest possible rating. The distributions for all twelve conditions are shown in Figure 2.

Table 1 shows the means for the different groups. For the ranking procedure, the point estimates are slightly above zero, but most of the 95% confidence intervals overlap with zero. For the scale procedure, those rating only one piece had the only negative point estimate. The rest are positive, but all their confidence intervals overlap with zero. The variances are higher for the ranking method and remain relatively stable across the number of items rated for both rating methods.

Inferential Statistics

A popular approach for modeling repeated measures data, particularly like these where people rated different art pieces, is as a cross-classified multilevel model, with random variables for both items and people. People rated between 1 and 10 items. Traditional repeated measures ANOVA have difficulty with different numbers of repeated measures, but multilevel models are well-suited for this (Goldstein, 2011; Wright & London, 2009). The `lmer` function from the R package **lme4** (Bates et al., 2015) will be used for these analyses. Following Luke (2017) and using the **lmerTest** (Kuznetsova et al., 2017) package, Satterthwaite approximations will be used for the degrees of freedom to calculate p-values of the fixed effects.

The results will be described in the order in which they were conducted. Variables were added to the model that predicted the standardized responses using just random variables for the participant and for the art piece. This baseline model did not include an

intercept so assumed the overall mean was zero, which because of how the variables were standardized is the mid-point of the scales and the value expected if there was no bias. The first variable added was to estimate the intercept. This test (RQ 3) was non-significant: $F(1, 9.13) = 0.57, p = .468$. This shows that there is no significant bias, overall, for the judgments. Next, the variable for which rating procedure was used (RQ 4) was entered and the means were not significantly different: $F(1, 663.58) = 2.68, p = .102$. It is worth mentioning the differences in variances observed in Table 1 are significantly different. Allowing different variances for the art pieces depending on the scale improved the fit: $\chi^2(2) = 60.98, p < .001$.

Next, a dummy variable for whether the participant rated all ten items or just a subset was entered into the model to test one of the study's main questions (RQ 1). This effect was non-significant: $F(1, 309.18) = 0.15, p = .695$. The interaction between this variable and the rating procedure used (RQ 6) was also non-significant: $F(1, 308.40) = 1.80, p = .181$.

The second main question (RQ 2) was whether adding a variable for whether the participant was rating just one item improved the fit of the model. It did: $F(1, 2726.22) = 5.85, p = .016$. From Table 1 the means when only make one rating are lower than when making multiple ratings. Given there are two critical research questions it is worth reporting the Holm adjusted p-value for this test: $p = .031$. The coefficient estimate is: -0.08, with a 95% (using the likelihood profile) confidence interval of (-0.14, -0.01). Overall the standard deviation of the standardized response variable is: 0.53, so this shift is -0.14 of a standard deviation. Using Cohen's (1992) terminology this is a small effect ($.2\sigma$ is what he calls a *small* effect), but it is still large enough to be of concern for some purposes. The interaction between this variable and the rating scale (RQ 7) was non-significant: $F(1, 2681.99) = 1.79, p = .181$.

The results thus far show that rating all ten was not significantly different than rating a subset, but that when rating just one item the scores were significantly lower. The next

research question (RQ 5) concerns whether the number of items to rate, from two to five, made a difference. Both linear and more flexible relationships were explored and none were significant. For the linear effect: $F(1, 314.19) = 1.16, p = .282$. The interaction between this linear effect and rating scale was also non-significant: $F(1, 1027.10) = 1.32, p = .250$.

Finally, there is concern about the generalizability of all observed effects in the social and behavioral sciences. While future research is necessary for testing these findings across situations, the design of this study allows for testing if age, education level, and/or gender moderate this effect. Age was entered as a cubic spline, with one knot at the median, into the model with just the rating scale variable and did not significantly improve the fit: $\chi^2(3) = 1.29, p = .731$. Neither the interaction with rating a subset versus rating all ten ($\chi^2(3) = 3.21, p = .360$) nor with rating one versus a multiple item subset ($\chi^2(3) = 3.70, p = .296$) reached statistically significance.

The education variable was recoded into four categories: some or finishing high school; some college or an associates degree; bachelors degree; and higher degree. This was entered as a four-category nominal variable. The main effect of education level was non-significant: $\chi^2(3) = 3.34, p = .342$. The interactions with rating all ten: $\chi^2(3) = 0.91, p = .824$ and rating one versus a multiple item subset, $\chi^2(3) = 0.58, p = .900$, were also non-significant. Similarly for the gender variable (only those self-reporting as male or female are used in these analyses), there was no main effect of gender ($\chi^2(1) = 0.00, p = .999$) nor interactions by rating a subset versus all ten ($\chi^2(1) = 0.01, p = .921$) or rating one versus a larger subset ($\chi^2(1) = 0.89, p = .345$). The extremely high p-value (and the corresponding low χ^2) was examined for the gender main effect. While the means are close (for females: 0.044, for males: 0.042), other approaches to estimate p (e.g., Kenward-Roger approximation) also yielded high values.

Conclusion

The research set out to answer an applied question that also has important consequences for the science of how people make judgments. In many contexts survey and market researchers would like to present people with only a subset of items. Sometimes this is out of necessity because of the number of items is very large and there are concerns the fatigue would adversely affect response quality if all items were asked. In these cases random subsets can be presented, but other times interest is only with a specific subset of some larger superset so this nonrandom subset is used. If just asking for judgments for a non-random subset it would not be possible to determine if people's ratings of this subset were different from their views of the superset, or if judging a subset tended to produce ratings that were in general too high or too long. This research provides some reassurance that people are unbiased when rating a subset of items. Using a fairly large sample and two rating procedures, participants rating subsets of 2, 3, 4, and 5 items were not significantly different from those rating all ten. There was a small (0.13σ) decrease when rating only one piece of art. From an applied perspective, while further research across multiple contexts would be useful, the current research should not dissuade anyone from having people rate only a subset of items, providing the subset is larger than one item.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for expeimenters: Design, innovation, and discovery* (2nd ed.). Wiley.

Bradley, M., & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, *21*, 167–184.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi: 10.1037/0033-2909.112.1.155

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8(3)*(e57410). doi: 10.1371/journal.pone.0057410

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, *23*, 1–74.

Goldstein, H. (2011). *Multilevel statistical models (4th ed.)*. Chichester, UK: Wiley.

Hess, S., Hensher, D. A., & Daly, A. (2012). Not bored yet: Revisiting respondent fatigue in stated choice experiments. *Transportation Research Part A: Policy and Practice*, *46*, 626–644. doi: 10.1016/j.tra.2011.11.008

Klein, M., D ulmer, 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*, 474–483. doi: 10.1093/ijpor/edh041

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). **lmerTest** package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*, 1494–1502. doi: 10.3758/s13428-016-0809-y

Mason, W., & Suri, S. (2012). Conducting behavioral researchh on Amazon's Mechanical Turk. *Behavioral Research Methods*, *44*, 1–23. doi: 10.3758/s13428-011-0124-6

Russell, P. A., & Gray, C. D. (1994). Ranking or rating? Some data and their implications for the measurement of evaluative response. *British Journal of Psychology*, *85*, 79–92. doi: 10.1111/j.2044-8295.1994.tb02509.x

Sheehan, K. B., & Pittman, M. (2016). *Amazon's Mechanical Turk for academics: The HIT handbook for social science research.* Irvine, CA: Melvin & Leigh.

Wright, D. B., & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, *62*, 439–456.

Table 1

Mean values of the standardized responses (mid-point of scale is zero, minimum and maximum are -1 and +1) for the different conditions along with their 95% confidence intervals found with parametric bootstrap (10,000 replications) using the defaults of the `confint.merMod` *function (Bates et al., 2015). The final row shows the mean of individual participants' variances (variances rather than standard deviations so that differences among conditions were not an artefact due to the number of items rated).*

	Number of Ratings					
	1	2	3	4	5	10
Ranks	0.03	0.10	0.11	0.07	0.03	0.03
95% CIs	(-0.05, 0.11)	(0.03, 0.18)	(0.03, 0.20)	(-0.00, 0.15)	(-0.04, 0.10)	(-0.02, 0.08)
Vars	*	0.35	0.22	0.30	0.34	0.34
Scale	-0.08	0.02	0.06	0.07	0.03	0.04
95% CIs	(-0.16, 0.01)	(-0.06, 0.09)	(-0.03, 0.14)	(-0.01, 0.14)	(-0.04, 0.10)	(-0.01, 0.10)
Vars	*	0.24	0.24	0.20	0.21	0.24

Note. * because each participant in these conditions had only one rating so their standard deviations were not calculated.

Figure 1. The ten pieces of art that were used for this study. (*courtesy of Soni Wright*)

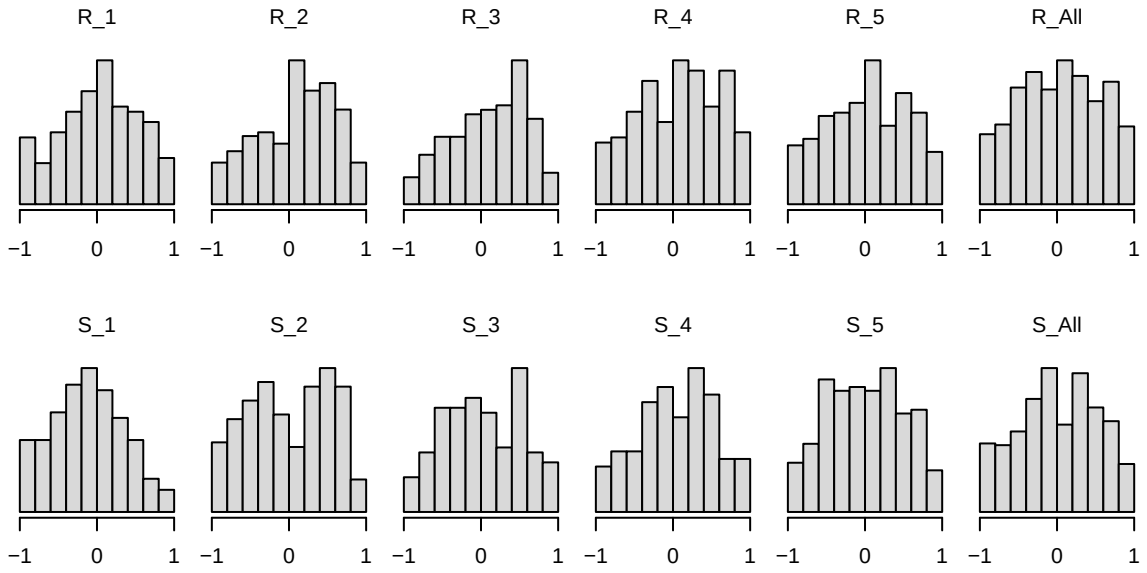

Figure 2. Distributions of standardized responses for the twelve conditions. The R (ranking) and S (scale) show which rating procedure was used and the numbers shows how many items were rated (or if all ten were rated).