Multipoint vs slider: a protocol for experiments

Venera Tomaselli, Giulio Giacomo Cantone

1. Introduction

Since the 1990s, in all fields involving survey tools aimed at collecting data from a sample of a target population, computer-assisted technologies of data recording replaced the old *paper-&-pen*. The speed of technological shift was not paired by methodological innovations.

Multipoint scales, indeed, are still among the most employed numerical (or semantic) supports for many variables in psychological, health, socio-economic research, and even in engineering (e.g., user experience design). With the spread of 'Big Data', an old issue in statistical measurement gained a new relevance. It can be shortly summarized: tons of Big Data from self-reports of taste and perception are recorded every day. While these data are reported through multipoint scales, almost all the relevant inferences are made through families of methods with parametric assumption, for example, one of the most notorious methodology to infer human preferences through analysis of similarity, *collaborative filtering* (Kluver, Ekstrand, and Konstan 2018).

The debate about the plausibility of an estimation of central value in ordinal variables (which is the core of the debate about parametric methods for analysis of 'ratings') is well summarised by Velleman and Wilkinson (1993). Kampen and Swyngedouw (2000) expanded the issue relating it the consequential debate about derivative measures of association and correlation among variables (~~also, see,~~ Agresti 2010). Tomaselli and Cantone (2020) highlighted a more recent issue in data analysis: when the number of items compared (e. g, a ranking) exceeds too much the categories of the supporting ordinal scale, the comparison is made impossible by the high amount of tie cases. Therefore, statistics constrained in the support scale (i.e., the median) are unfeasible to index distributions from very large samples, or populations. This problem of ranking statistics could be interpreted as an extreme case of 'ceiling effect' (Austin and Brunner 2003).

Slider scales, which are technological advancements not previously available on paper-&-pen survey but now enhanced by surveying with web tools, can overcome the issues of ordinal scales. A slider scale ('slider') is a bar representing a visually continuous segment of numerical points through 1 to *m* (sometimes through 0 to *m*, or to -*m* to *m*). While the number of points is finite, for any analytical purpose this measurement is considered continuous and not ordinal, therefore *m* should not be a small number. A very common case is *m* = 100.

The respondent moves an indicator ('it slides') among the values in the bar. If the bar is drawn on a paper, as in the case for Visual Analogue scales (VAS), the respondent can only appoint a mark on the bar. The estimate of VAS may be considered continuous, and more accurate than multipoint scales (Voutilainen *et al*. 2016), but the value would be technically harder to record. For years the absence of proper computing, visualizing, and recording technologies impacted the developments of statistical science. Could multipoint and Likert scales be reputed obsolete because they were designed for *paper-&-pen* data collection? Results from Fryer and Nakao (2020) validate this thesis, while a web experiment by Funke (2015) criticizes sliders. Other results (see, Roster, Lucianetti, and Albaum 2015; Bosch *et al*. 2018) bring further arguments on the evaluation of sliders, in particular reporting a longer time of completion of tasks. A comprehensive review of the debate is provided by Chyung *et al*. (2018).

Matejka *et al*. (2016) performed an experiment testing the accuracy of sliders compared to a Likert scale and on the impact of marks with percentages ('ticks') on the bar of sliders. Participants

Venera Tomaselli, University of Catania, Italy, venera.tomaselli@unict.it, 0000-0002-2287-7343
Giulio Giacomo Cantone, University of Catania, Italy, giulio.cantone@phd.unict.it, 0000-0001-7149-5213

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Venera Tomaselli, Giulio Giacomo Cantone, *Multipoint vs slider: a protocol for experiments*, pp. 91-96, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-304-8.19, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci, *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the opening conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-304-8 (PDF), DOI 10.36253/978-88-5518-304-8

(n = 2000) were recruited through *Amazon's service Mechanical Turk*. Participants were asked to estimate the blackness of a shade of grey through sliders or Likerts. Results show that sliders without ticks have better performances in both accuracy of the judgements and bias reduction. Even if authors do not mention it directly bias observed in their results is coherent with the psychological phenomenon of *heaping*, a connection rarely mentioned (an exception: Couper *et al*. 2006).

To monitor heaping effects is important because, while in scales with ticks heaping is due to psychological attachment, there is evidence that heaping is also related to fabricated data in data collections (Finn and Ranchos 2015).

2. Experimental protocol

The sample of respondents is recruited through a web open procedure, like the aforementioned Mechanical Turk. The survey tool is therefore a website. The data collection process is segmented in 3 phases. After completion of 1st phase, a new record is added to a connected database while 2nd and 3rd phases add more data to the record.

In the 1st phase participants are randomly assigned to two random treatment groups. Both the groups are assigned to a task or 'trial': they have to estimate the colour of a square. This trial is repeated for 10 times. The treatment difference among the two groups is that the control group has to estimate the colour through a 0-10 multipoint scale, while the experimental group has to estimate it through a 0-100 slider bar.

As showed in Matejka *et al*. (2016), estimation of shades of colours through a sequence of trials is among the best for objective evaluation of measurement tools (i.e., scales). Instead of presenting to respondents 50 fixed shades of grey squares, we propose a random generator of a shades of Red and Blue. A square of Yellow is superimposed with an opacity randomly distributed between 0% and 10%. Therefore, any randomly coloured square is a realization of the combination of: (i) a randomly generated parameter ξ of shade, uniformly distributed between 0% (full Red) and 100% (full Blue) and (ii) a randomly generated parameter ζ of noise, uniformly distributed between 0% and 10%.

In the 1st phase participants are requested to estimate only shade, with opacity being a possible factor of controlled noise. In the original experiment of Matejka *et al*. there was no mechanism to control noise in the estimation process, even if authors accounted that differences in participants' devices should have been factors of noise out of experimental control. Another difference from Matejka *et al*. is that participants should be free to refuse to complete any trial. The default option in a Likert scale, signalled through a button under (not adjacent) the multipoint scale, is 'no answer'. The best equivalent to let "no answer" in a slider would be setting invisible the indicator on the bar before interaction to it, providing a button 'no answer' to remove it again. This does not push a heaping bias inflation towards initial positions of the indicator (Liu and Conrad 2018). In this case, if the respondent avoids interacting with the slider, a 'no answer' is recorded.

The software must record not only the final choice of the participants but also every single interaction with the tool, tracing their decisional process. Continuous sliders are very well suited for this tracing because there is a large support of values to pick on.

When a participant completes 1st phase, data recorded is: (i) random generated shade parameter ξ for the 10 trials; (ii) random generated opacity parameter ζ for the 10 trials; (iii) participant's estimations x for the 10 shades; (iv) time of completion t_x for each of the 10 trials; (v) number of clicks k_x for each of the 10 trials.

In the 2nd phase participants are asked to report their *taste*-response of 10 well-known leisure products through the scale (to rate) of their treatment groups in 1st phase. When the participant completes the 2nd phase, further information can be added to the record: (vi) participant's rating r for each of the 10 products; (vii) time of completion t_r for each of the 10 ratings; (viii) number of clicks k_r for each of the 10 trials. If the rating process is interrupted, no data is added to the record.

In the 3rd phase standard demographic variables are collected from participants, whereas they provide consent.

3. Methods of data analysis

Heaping is a relevant bias in applied statistical studies on scales of measurement. Even if they do not mention it directly, the statistic adopted in Matejka *et al.* (2016) to measure heaping is a normalised score of the mean deviation from the expected difference of observed frequency among adjacent values:

$$\sqrt{\frac{\sum\left((n_x - n_{x-1}) - \frac{\sum n_x - n_{x-1}}{|M|}\right)^2}{|M|}} \qquad (1)$$

where |*M*| is the cardinality of the support, *x* is the observed value from the M scale and *n* is the absolute frequency associated to *x*[1]. Matejka *et al.* reported a score of heaping ~ 2 (± 0.1 at CI 95%) for sliders, while the introduction of 'ticks' that imitate multipoint scales in the slider significantly increases the heaping bias (Fig 1, see "no ticks"). The relation is not linear to the number of ticks.

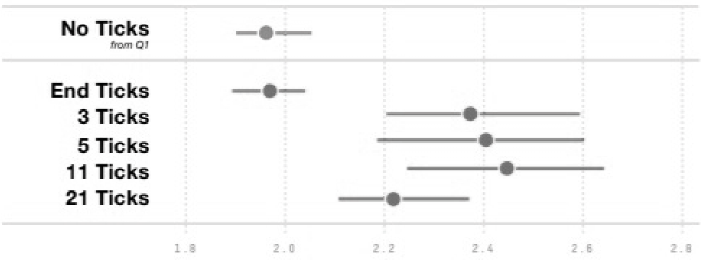

Figure 1. Mean heaping scores for varying number of tick marks. Error bars show 95% CIs. (Matejka *et al.*, p. 5)

We make the hypothesis that control group (multipoint) induces more heaping than experimental group (sliders).

Since values (*x* for estimates of shades, *r* for ratings on products) from sliders and multipoint scales are constrained in a finite support, they can be normalised into a [0,1] interval. The distribution of errors $\xi - x$ is the main statistic and is assumed to be normally distributed. A Shapiro-Wilk test is performed on the sample of $\xi - x$ values of all the trials *per* group to confirm this assumption. Since noise factors ζ are all sampled from the same population, we expect no significant difference in the distribution of values. This assumption is tested through a Kolmogorov-Smirnov test. If violated, $\xi - x$ values will be controlled *per* ζ. Times of completion t_x are assumed to be normally distributed. This assumption is tested through a Shapiro-Wilk test.

Null hypotheses on the *objective task* of shade estimation with random noise are:

i. sliders induce a distribution of mean absolute errors (MAE) from randomised parameters over the 10 trials which is not superior to multipoint scales' MAE.

Absolute errors $| \xi - x |$ are never assumed to be distributed normally: if $\xi - x$ values were

[1] Is there an implicit consensus of statistical science on this measure? Roberts and Brewer (2001) provide 2 different approaches to measure heaping: (i) H_1 is technically only a minor improvement over (1) while (ii) C_2 is based on the probability to observe local modes. The (ii) approach raises issues on the confidence threshold to assert that an observed local mode is *likely a true* local mode and not a local noise. For a modern approach to heaping models, see Zinn and Würbach (2015)

normally distributed, then their absolute values would be distributed as half-normal distribution (Folded Normal). Given the structure of the hypothesis, a non-parametric 1-tailed test (i.e., Mann-Whitney test) on the samples of participants' MAE in the two groups (a MAE *per* participant) seems suited to check the hypothesis.

ii. sliders induce less variance and not superior skewness than multipoint scales. If $\xi - x$ values of all the trials *per* group are normally distributed, the exact 1-tailed Fisher's test of variance (*F*-test) is suited to check the hypothesis on variance. If the errors are not normally distributed, the non-parametric alternative will be 1-tailed Levene's test. The simpler test to check if treatment variable induces a systemic error in objective estimation is the test of signs of $\xi - x$. A significant difference from null hypothesis of sum of signs equal to 0 for both groups will need to be commented.

iii. sliders induce a not superior t_x than multipoint scales. If time values of all the trials *per* group are normally distributed, a 1-tailed *z*-test of means will check the hypothesis. If time values are not normally distributed the non-parametric alternative is 1-tailed Mann-Whitney test.

Correlations between degrees of controlled noise ζ, errors $\xi - x$, times of completion t_x, and clicks k_x are graphically represented through scatterplots and visualised through a generalised model if the fit is sufficiently good. The effect of noise on $\xi - x$ is supposed to be non-linear and possibly not even symmetrical around the value of $\xi - x = 0$, although it can be symmetrical around a different value. Noise can similarly affect t_x and k_x, too.

Does the same structure of hypotheses A, B, and C hold for measures collected in 2nd phase? Since the 10 leisure products have to be chosen among well-known, a prior value ρ of expected taste can be elicited through an expected value computed from rating statistics of online rating platforms. Although arguably biased for both small and large samples (Askalidis, Kim, and Malthouse 2017), these priors are likely the most reliable predictors of expected *taste* at least from a population of subjects very interested in the product category[2].

Even accounting for aforementioned biases, the statistic $r - \rho$ can be interpreted as a *deviation* of biased raters *vs.* randomised raters. Even if $| r - \rho |$ and $| \xi - x |$ are technically the same operation of *distance*, their arguments are conceptually distinct, as reflected through the order of minuends and in the semantic difference between an *error* (there is always a true parameter ξ) and a *deviation* (two procedures to evaluate the same *evaluando*). As a consequence, the hypotheses on $r - \rho$ cannot be 1-tailed. However, although *tastes* are not *objective*, hypotheses on the differences in values, variances, and skewness among groups can still be asserted.

Moreover, means of $r - \rho$ values can be both correlated and compared to paired (intra-participant) means of $\xi - x$ values (controlled on ζ). Correlating and comparing times of completions (t_x with t_r) and clicks (k_x with k_r) is even less ambiguous since they measure both the same physical quantities. Differences and ratios between the two phases can be compared *per* group, too.

Finally, whereas the sample sizes on demographics collected in 3rd phase support it, associations between demographic variables to aforementioned statistics can be asserted as a control procedure but no causal explanation emerges from literature about trials on the colour perception.

4. Conclusions

While this protocol partly replicates the experiment of Matejka *et al.* (2016), we propose some relevant improvements to define a general experimental protocol for data collection and analysis on web-tool of human perception and tastes:

[2] For example, the rating platform *Letterboxd* reports that the movie *The Godfather* (directed by Francis F. Coppola, released in 1972) received more than 300,000 ratings from all over-the-world raters. According to Lorenz (2006) even in presence of local peaks, the best models to represent movies have only one location parameter, which the author interprets as an "evidence of universality in processes of continuous opinion dynamics about taste" (p. 251).

- we generalise the structure of hypotheses that tests the statistical efficiency of the measurement tool through web trials. While hypotheses A (location) and C (duration) were already well-covered in literature, hypothesis B (variance) is often neglected. The definition of statistical assumptions makes explicit some elements of potential fragility of previous literature on the topic of evaluation of measurement tools for social sciences, i.e., to our knowledge no research on sliders mentioned the potential need of non-parametric tests for variance of errors or deviations.
- the previous issue is likely the consequence of a general under-recognition of research of heaping bias. Matejka *et al*. (2016) did not acknowledge literature on heaping. We connected their empirical work to the at-state-of-art mathematical alternatives for measurement of heaping bias. We also re-wrote (1) in a less ambiguous and friendlier formalism for statisticians and psychometricians.
- we see improvements in the experimental procedures, since we introduced a noise parameter ζ that affects the coloured square inducing visual opacity. This inclusion reproduces better extra-experimental situations of perception.
- the inclusion of a data collection on tastes in the 2nd phase provides not only a better assessment on scales' performance but it could also highlight insights on the relationships between *perception* and *taste*. Of course, we assumed that an experiment focuses on a particular taste for *something* (e.g., movies are convenient) but further experiments could pair perceptions and ratings on different objects (arts, languages, etc.).

The major rationale to adopt sliders has sprouted from the theoretical debates mentioned in Section 1, so far. For applied research, even in absence of evidence of remarkable improvements (see, hypotheses A, B, and C in Section 3) in the reduction of coarseness in data, inaccuracies of self-report, and biases through adoption of sliders, the evidence that sliders reduce scale-induced heaping (Figure 1) is extremely insightful. Better measurement scales can minimize the confounding effect in those research programmes aimed to investigate data fabrication (i.e., fraud reports) through tests on heaping.

References

Agresti A. (2010). *Analysis of Ordinal Categorical Data*, Wiley, Hoboken, (NJ).

Askalidis, G., Kim, S.J., Malthouse, E.C. (2017). Understanding and overcoming biases in online review systems. *Decision Support Systems*, **97**, pp. 23-30.

Austin, P.C., Brunner, L.J. (2003). Type I error inflation in the presence of a ceiling effect. *The American Statistician*, **57**(2), pp. 97-104.

Bosch, O.J., Revilla, M., DeCastellarnau, A., Weber, W. (2018). Measurement reliability, validity, and quality of slider versus radio button scales in an online probability-based panel in Norway. *Social Science Computer Review*, doi:10.1177/0894439317750089.

Chyung, S.Y.Y., Swanson, I., Roberts, K., Hankinson A. (2018). Evidence-based survey design: The use of continuous rating scales in surveys, *Performance Improvement*, **57**(5), 38-48.

Couper, M.P., Tourangeau, R., Conrad, F.G., Singer, E. (2006). Evaluating the effectiveness of visual analog scales. *Social Science Computer Review*, **24**(2), pp. 227-245.

Finn A., Ranchhod V. (2015). Genuine fakes: The prevalence and implications of data fabrication in a large South African survey. *The World Bank Economic Review*. doi:10.1093/wber/lhv054.

Fryer, L.K., Nakao, K. (2020). The future of survey self-report: An experiment contrasting Likert, VAS, slide, and swipe touch interfaces. *Frontline Learning Research*, **8**(3), pp. 10-25.

Funke, F. (2015) A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales, *Social Science Computer Review,* **34**(2), pp. 244-254.

Kampen, J., Swyngedouw, M. (2000). The ordinal controversy revisited. *Quality & Quantity,* **34,** pp. 87-102.

Kluver, D., Ekstrand, M. D., Konstan, J. A. (2018). Rating-based collaborative filtering: algorithms and evaluation. In *Social Information Access,* eds. P. Brusilovsky and D. He, Springer, Charm, (SW), pp. 344-390.

Liu M., Conrad, F.G. (2018). Where should I start? On default values for slider questions in web surveys, *Social Science Computer Review*. doi:10.1177/0894439318755336.

Lorenz, J. (2006). Universality in movie rating distributions. *The European Physical Journal B.* **71**, pp. 251-258.

Matejka, J., Glueck, M., Grossman, T., Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. doi: 10.1145/2858036.2858063.

Roberts, J.M., Brewer, D.D. (2001). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, **28**(7), pp. 887-896.

Roster, C.A., Lucianetti L., Albaum, G. (2015). Exploring slider vs. categorical response formats in web-based surveys, *Journal of Research Practice*, **11**(1), Article D1. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/509/413.

Tomaselli, V., Cantone, G.G. (2020). Evaluating Rank-Coherence of Crowd Rating in Customer Satisfaction. *Social Indicators Research*. doi: 10.1007/s11205-020-02581-8.

Velleman, P.F., Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, **47**(1), pp. 65-72.

Voutilainen, A., Pitkäaho, T., Kvist, T., Vehviläinen-Julkunen, K. (2016). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing*, **72**(4), pp. 946-957.

Zinn, S., Würbach, A. (2015). A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics*, **43**(4), pp. 682-703.