Reducing nonresponse to open-ended questions in mobile surveys of instructional quality. A Sequential Bayes Factor Analysis of two experiments.

Samuel Merk[1], Augustin Kelava[1], & Thorsten Bohl[1]

[1] Eberhard Karls Universität Tübingen

Author Note

Samuel Merk, Faculty of Economics and Social Sciences, University of Tübingen; Augustin Kelava, Faculty of Economics and Social Sciences, University of Tübingen; Thorsten Bohl, Faculty of Economics and Social Sciences, University of Tübingen;

Correspondence concerning this article should be addressed to Samuel Merk, Muenzgasse 22. E-mail: samuel.merk@uni-tuebingen.de

Abstract

Mobile surveys of instructional quality are a frequently used method to . . . . Whereby answers to closed answers . . . answers to open-ended questions function as . . . . To exploit this potential, teachers need narrative answers of high quality, which are traditionally difficult to access .. high rates nonresponse . . . We report on two large experiments aimed at reducing nonresponsebias

Keywords: nonresponse; open-ended questions; survey; instructional quality; feedback

Word count: X

Reducing nonresponse to open-ended questions in mobile surveys of instructional quality. A Sequential Bayes Factor Analysis of two experiments.

Introduction

OEQs educational contexts

Assessing OEQs with mobile surveys in educational contexts

. . . lead us to the following research questions:

> How can nonresponses to open-ended questions in mobile surveys on instructional quality be reduced by the placement of the question at the beginning of the questionnaire, and the addition of an motivating introduction to the question stem?

We aimed to provide answers to these research questions by conducting two experiments (Study 1 and Study 2). Both were conducted in the context of private tutoring. Students, who took private tutor lessons for one week in math, rated the instructional quality of this course using an adapted version (Merk, Poindl, & Bohl, n.d.) of the SEEQ (Marsh, 1982). Additional to this likertype items, students were asked four open-ended questions about the strength and weakness of the teacher (OEQ 1 & 2) and the course provider (OEQ 2 & 3).

Sequential Sampling

Both studies which we conducted, were field studies and constitute a randomized manipulation of a large, ongoing, “in production” feedback system. This randomized manipulation contained conditions which were assumed to improve the feedback system by gathering better data (less nonresponse, longer narrative answers). Hence, it is ethically implied to shift the whole feedback system towards the best experimental condition, as soon as there is enough evidence for this conditions superiority.

But despite the plausibility of this strategy it is very challenging to implement it using traditional research designs (e.g. randomized controlled trial) and frequentist statistics. Remember that in a best practice version the Neyman-Pearson procedure requires to *1)* define a minimal meaningful effect size, *2)* define the tolerated false-positive rate α, *3)* run an a priori power analysis based on a data analysis model and the results of the two previous steps, *4)* run the study and *5)* compute the p-value and reject H_0 if $p < \alpha$ (Cohen, 1988). Henceforward, this procedure is called Null-Hypothesis Significance Testing with a priori Power Analysis (NHST-PA; F. Schönbrodt et al., 2017). The challenge of applying this procedure to the problem at hand, lies in the fact, that the NHST-PA is a so called fixed-n sampling design (Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019), meaning that the sample size has to be predetermined and a optional stopping or continuation of data collection in dependence of the intermediate results invalidates the interpretation of resulting p-values or the confidence intervals (Lai, Lavori, & Shih, 2012).

However statisticians have developed several extensions of the NHST-PA which allows to achieve correct p-values while looking repeatedly into the data and deciding about further data collection in dependece to the interim results (sequential design). The most common of such sequential designs is the so called group sequential (GS) design (Lai et al., 2012). GS allows for a controlled over all Type I error rate with interim tests (e.g. $n_1 = 250$, $n_2 = 500$, $n_3 = 750$) and a final test (e.g. $n_{fin} = 1000$) even when researchers optionally stop the data collection after a interim test, depending on the results. However, planning a GS, requires researchers to prespecify all interim tests before the data collection starts (Proschan, Lan, & Wittes, 2006).

For the current problem GS are not suitable, as the data occurs in blocks (per course) which sizes are not susceptible to reseachers. Hence, we decided to use a Sequential Bayes

Factor Design (SBF, Schönbrodt et al., 2017). The SBF is based on Bayesian Hypothesis Testing and uses the Bayes Factors (BFs) as measure of evidence. BFs are formally defined as $BF_{10} = \frac{p(D|H_0)}{p(D|H_A)}$ and hence express to what extend given data is more compatible to H_0 or H_A. The SBF ca be described as a procedure with the following steps: *1)* A priori definition of a treshold of evidence (e.g. $BF_{10} = 10$). *2)* Description of the plausibility of effect sizes under H_A (prior distribution of effects). *3)* Data collection for a minimum sample size. *4)* Alternating calculation of the BF and arbitrary increase of sample size. The last step means, that using SBF researchers can do sequential sampling in a really flexible way. E.g. in a lab study it may make sense, to calculate the BF after each participant, whereas in the current studies it seems to be appropriate to calculate the BF after each evaluated course.

Overall, SBF is a flexible method, which allows researchers to monitor the emerging evidence while data is accumulating, and optionally stop the experiment when the evidence exeeds a predefined level. Thereby no Type I error inflation occurs (Rouder, 2014) and no predefinition of effect sizes is necessary.

General Method

Two studies with SBF . . . one exploratory and one extended replication . . .

Study 1

Design

In Study 1, we investigated the research questions using a 2x2 between-person design: The first experimental factor was the *position* of the OEQ within the questionnaire with the two steps *at the beginning* of the questionnaire and *at the end* of the questionnaire. The second factor was the presence of an additional *motivating introduction to the question stem* with the steps *additional motivating introduction* and *no introduction.*

Procedure and Materials

After the fourth day of their five days lasting course, students were asked to fill out a questionnaire to give the teacher and the school feedback about the lessons. The questionnaire contained the items of the SEEQ and two OEQ about instructional quality (see section Measuremets). Students used their own mobile phones for the survey and were given course specific short link which resulted in a customized survey which mentioned the name of the respective teacher in the introduction and some items (e.g. *TEACHERNAME encourages students to participate in class discussions*).

Participants

Measurements

SEEQ. We assessed the instructional quality using an adapted version of the Students' Evaluations of Educational Quality questionnaire (SEEQ, Marsh, 1982). . . .

```
## $\chi^2$ = 1627.791, _df_ = 137, CFI = 0.922, TLI = 0.902, RMSEA = 0.061, SRMR = 0.04
```

Open-ended Questions. The two OEQ of the survey were *What has TEACHERNAME done especially well?* and *What could TEACHERNAME do better in future?*. Over all questions and experimental conditions 73.87% of students gave answers, which were at average

Statistical analysis

Results

Study 2

Design and Procedure

Both

Participants

Measurement

Statistical analysis

Results

Discussion

References

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New: Lawrence Erlbaum.

Lai, T. L., Lavori, P. W., & Shih, M.-C. (2012). Adaptive Trial Designs. *Annual Review of Pharmacology and Toxicology*, *52*(1), 101–110. doi:10.1146/annurev-pharmtox-010611-134504

Marsh, H. W. (1982). Seeq: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*(1), 77–95. doi:10.1111/j.2044-8279.1982.tb02505.x

Merk, S., Poindl, S., & Bohl, T. (n.d.). Wie sollten Rückmeldungen von quantitativ erfasstem Schülerfeedback (nicht) gestaltet werden? Wahrgenommene Informativität und Interpretationssicherheit von quantitativen Rückmeldungen zur Unterrichtsqualität. *Unterrichtswissenschaft.* doi:10.1007/s42010-019-00048-5

Proschan, M. A., Lan, K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials. A unified approach.* Springer Science & Business Media.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi:10.3758/s13423-014-0595-4

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. doi:10.1037/met0000061

Schönbrodt, F., Gollwitzer, M., & Abele-Brehm, A. (2017). Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien. *Psychologische Rundschau*, *68*(1), 20–35. doi:10.1026/0033-3042/a000341

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*. doi:10.3758/s13428-018-01189-8