Filtering Survey Responses from Crowdsourcing Platforms: Current Heuristics and Alternative Approaches

Completed Research Paper

Lennard Schmidt
HHL Leipzig Graduate School
of Management
Jahnallee 59, 04109 Leipzig
lennard.schmidt@hhl.de

Florian Dost
Alliance Manchester
Business School
Booth St W, Manchester M15 6PB
florian.dost@manchester.ac.uk

Erik Maier
HHL Leipzig Graduate School
of Management
Jahnallee 59, 04109 Leipzig
erik.maier@hhl.de

Abstract

Information Systems research continues to rely on survey participants from crowdsourcing platforms (e.g., Amazon MTurk). Satisficing behavior of these survey participants may reduce attention and threaten validity. To address this, the current research paradigm mandates excluding participants through filtering heuristics (e.g., time, instructional manipulation checks). Yet, both the selection of the filter and the filtering threshold are not standardized. This flexibility may lead to suboptimal filtering and potentially "p-hacking", as researchers can pick the most "successful" filter. This research is the first to tests a comprehensive set of established and new filters against key metrics (validity, reliability, effect size, power). Additionally, we introduce a multivariate machine learning approach to identify inattentive participants. We find that while filtering heuristics require high filter levels (33% or 66% of participants), machine learning filters are often superior, especially at lower filter levels. Their "black box" character may also help prevent strategic filtering.

Keywords: Survey Research, Filtering, Amazon MTurk, Reliability, Validity, Effect Size

Introduction

Online survey studies with respondents from crowdsourcing platforms such as Amazon Mechanical Turk have become a cornerstone of behavioral research (Mason and Suri 2012). In Information Systems research, for example, the number of papers at the annual International Conference of Information Systems referring to "mechanical turk" or "MTurk" as their respondent pool has steadily increased, from 13 in 2014, to 25 in 2018. While some authors evaluate the use of MTurk as participant pool critically (Cheung et al. 2017; Wessling et al. 2017), other evaluations see response qualities on par with other frequently used samples (such as students [Hauser and Schwarz 2016] or other survey pools [Owens and Hawkins 2019]). However, while studies on MTurk or similar platforms might thus offer quick and inexpensive results, satisficing behaviors of semi-professional MTurkers (Schmidt 2015) who may inattentively rush through the surveys

(Cheung et al. 2017), potentially using answering aids or even bots (Stokel-Walker 2018), present a continuing threat to reliable and valid results. For a survey researcher, this necessitates respondent filtering procedures.

Filtering survey respondents introduces two important researcher decisions: First, a researcher needs to decide which filter to use, and second, how sensitive or strictly to apply it (i.e., select a threshold). The first decision needs to consider transparent, objective, and replicable criteria to check a respondent's answers' validity and reliability. The second decision needs to consider the trade-off between smaller errors in the observable measures, correlations, and effects on the one hand, but smaller samples and hence lower power to detect such effects on the other hand. Ideally, a transparent and replicable algorithm balances metrics and checks for both, answer validity and reliability. In principle, such algorithm results would rank respondents in the order that ensures maximal marginal gain in survey quality per respondent that is filtered. These gains, for example in effect sizes, could then be contrasted with the loss in power to reliably detect an effect as more and more respondents get filtered from the sample. Surprisingly, no research on online survey studies has explored filtering procedures that comprehensively link criteria of reliability and validity, compared filtering procedures per numbers of respondents filtered, or considered sampling considerations such as statistical power.

Extant research on filtering procedures has often focused on detecting respondent inattention, arguing that such "subject inattention" threatens the validity of the results (Cheung et al. 2017, p. 356). Common procedures therefore involve attention checks, such as instructional manipulations checks (IMCs), to test if a participant pays close attention (Paas et al. 2018; Paolacci et al. 2010). Other filtering procedures assess respondent attention through differences in survey completion times (Warkentin et al. 2017). For example, when the average time per word as a measure for reading speed exceeds a certain threshold, a respondent gets filtered (Huang 2014). In consequence, in current Information System research, the actual filtering decision takes the form of ad hoc heuristics selected by the researcher, such as filtering all respondents who fail a number of IMCs (Barlow et al. 2018), do not complete certain items (Ahn et al. 2018), whose answers do not fit theory (Clemons et al. 2016), or a combination of multiple measures (Warkentin et al. 2017). Importantly, filters are not only applied to select respondents from online survey pools, such as MTurk, but also to select participants among actual users of a service (Ahn et al. 2018; Barlow et al. 2018). Besides a selection of the filter, researchers also have to set the thresholds for these heuristics filters (e.g., not all IMCs correct, more than two standard deviations from average completion time). However, whether and at which threshold such a heuristic appropriately distinguishes between attentive and inattentive participants to increase validity, and how the filtered number of respondents actually affects the reliability of the measures or the power to detect an effect, is not discussed. This is particularly apparent in the few studies that compare different filtering procedures, such as different types of IMCs, but fail to consider results based on similar remaining sample sizes (e.g., Peer et al. 2014).

The present research addresses this gap, aiming to create transparency about the effect of different filters and to provide guidelines for filtering participants on MTurk and other participant pools (i.e., also representative samples). First, a survey study is set up to collect data on several common filtering heuristics simultaneously. These include different IMCs and factual attention checks, a reading speed task and tracking reading speeds at the survey, page and question level. These validity-based checks and metrics are complemented with metrics for reliability, measuring answering consistency (e.g. Cronbach's alpha between scale items of a construct), or test-retest correlations. Second, using this comprehensive set of available metrics, several filtering procedures are compared against each other. In this, the filtering consequences for effect sizes are compared, using manipulations for two well-established psychological effects, which are commonly used to assess data quality by means of the measured effect sizes (Paolacci et al. 2010). Third, we introduce a machine learning "black box" approach which increases the filtering quality on all key metrics (validity, reliability, etc.) and opens up opportunities for further reducing researcher degrees of freedom – a common driver of so-called "p-hacking" (Simonsohn et al. 2014). Importantly, our analysis does not intend to provide researchers with a toolbox to filter for a desired p-value, but rather aims to raise awareness for the effect of different filtering techniques, which might ultimately result in stricter standards of the research community (e.g., in terms of appropriate filters).

Our findings offer several contributions: Comparing different filtering procedures reveals that not all procedures improve assessed answer validity beyond random filtering. Those procedures that actually do, improve answer validity on different trajectories, i.e., sooner, with fewer filtered respondents, or later, with more filtered respondents. In particular, attention tests and filters based on overall time increase the validity of the results. In contrast, filters by page (e.g., either as total time, reading speed or variance) perform badly. Additionally, many filters only improve results at higher filter levels. Furthermore, the filtering procedures differ considerably regarding the attainable psychological effect sizes per filtered respondent, and the associated loss in statistical power. These findings imply that differences between filtering procedures exist, and that they occur gradually at different thresholds (i.e., number of filtered respondents) – a finding that is new to survey researchers who previously tended to set filtering sensitivities ad hoc. We, therefore, introduce a machine learning filter that incorporates not only one heuristic (e.g., completion time) but multiple variables. This more complex procedure outperforms the simple heuristics. We believe these findings encourage a discussion about the present methodological paradigm in Information Systems research to use different heuristic filters to identify inattentive participants.

Background

Filtering respondents from survey studies is a controversial topic. Social science survey studies in general have come under scrutiny over discussions about the low replicability of prominent findings and effects (e.g., Open Science Collaboration 2015). One key observation is the role of small samples and low power research designs that may allow an artificially high number of chance results to pass as publishable findings (Button et al. 2013; Simmons et al. 2011). Filtering specific subjects, cases, or respondents not only directly reduces sample size and power, it also presents a researcher’s degree of freedom, which may willingly or unwillingly foster “p-hacking” (Simonsohn et al. 2014) through filtering out respondents with an answering pattern that goes counter to an intended hypothesis. In this light, filtering needs to follow replicable, transparent and objective procedures. Furthermore, even for a procedure that meets these requirements, setting a filtering threshold is a researcher degree of freedom that could affect measured effects and power, and in the worst case cause bias in the results.

On the other hand, not filtering potentially biased responses increases error in the results and the need for larger samples in case these errors are random; in case the errors follow from systematic unwanted respondent behaviors, not filtering may severely bias the results (Barber et al. 2013). In online survey studies on crowdsourcing platforms, several unwanted behaviors can be expected (van Herk et al. 2004). Respondents often receive a fixed compensation for their efforts, such as course credit for students or payments for MTurkers. As a result, respondents have an incentive to reduce their efforts, resulting in several forms of satisficing behavior (Krosnick 1991; Oppenheimer et al. 2009). For example, respondents speed through the survey (Malhotra 2008), or click Likert scales in a straight line (Zhang and Conrad 2014), or learn to mimic seemingly plausible responses patterns. All of these behaviors limit the validity and reliability of the survey studies. When administering attentiveness checks, up to 46% (Oppenheimer et al. 2009) of the respondents from crowdsourcing platforms or students samples frequently fail to respond correctly, suggesting that potentially biasing behaviors cannot easily be disregarded and that biases would not dilute quickly as sample sizes are increased. Taken together, filtering done right is necessary for valid and reliable survey research. Filtering procedures need to assess criteria that are reasonably linked to unwanted respondent behaviors, while not being directly linked to the investigated effect.

A commonly used filter procedure uses survey completion time. For example, respondents who completed the survey quicker than an absolute threshold time were excluded from the analyses (e.g., Warkentin et al. 2017). The idea is that inattentive respondents speed through a survey, and quick completion times would not have allowed proper reading of instructions and questions. When accounting for variable survey length, relative measures, such as time per word, have been applied to filter inattentive participants (Huang 2014). However, respondents’ reading speed differs widely by respondent age (Bui et al. 2015) or between different professional groups (e.g., fast-reading college professors: Kershner 1964). To address individual differences in reading speeds and allow individual-level reading speed assessments, specific reading tasks can be used (e.g., Moore 2015). Still, in the following actual survey, reading speeds vary widely, within an individual respondent, from page to page (Huang et al. 2012). As a result, using threshold completion times, or reading speeds by page or question, or reading speed tests could be used as readily available filtering procedures, but it is unclear how they affect the validity and reliability of the responses, or how sensitive the respective thresholds needs to be selected to obtain the maximum gain in survey quality with minimal loss in statistical power.

Other approaches assess respondents' attention directly with instructional manipulation checks (IMCs: Paas et al. 2018; Paolacci et al. 2010) or factual manipulation checks (Downs et al. 2010). In the context of crowdsourcing platforms, however, such tests may offer insufficient quality control: MTurkers for instance, are often experienced survey takers (Schmidt 2015), conditioned to spot IMCs (Hauser and Schwarz 2016) and knowledgeable of common IMCs (Goodman and Paolacci 2017). Information about the presence of IMCs and how to pass them is sometimes shared in forums (Chandler et al. 2014; Wessling et al. 2017). Such professionalism on IMC taking may even affect platform reputation measures, such as successful survey completion rates in the past, which are also used by researchers to filter respondents (Peer et al. 2014). Finally, especially when using several IMCs, it remains unclear whether it is necessary to filter all respondents who failed any test (e.g., Barlow et al. 2018; Teubner and Flath 2019), which often would halve the available sample size.

Summing up the literature on filtering procedures, it is apparent that no consensus on the right filtering procedure exists and that the consequences of the most common filters (e.g., time, IMCs) are mostly unknown. Therefore, it seems necessary to compare existing filters. Additionally, more complex, multivariate filters may offer superior filtering characteristics, as it seems unlikely that a single measure captures all forms of respondent inattentiveness. The "black box" character of such approaches – namely machine learning based on all available variables – may also help to reduce researcher degrees of freedom, thus attempting to balance the needs for reliable and valid responses with the dangers of p-hacking.

Empirical Investigation

Research Methodology

Setting and participants: We asked crowdworkers on Amazon MTurk to participate in a survey on online privacy concerns; all components of the survey related to this topic. Participants received $1 in compensation for the 10 minute survey, which equals the frequently suggested fair minimum pay (Peer et al. 2014) and renders the survey eligible for internal approval in the MTurk community ("HITs worth turking for", Schmidt 2015). Only high reputation workers (at least 95% approval rate) with a minimum of 500 approved tasks were able to participate, as suggested in extant research (Peer et al. 2014). 198 participants completed the questionnaire (mean age: 35.75, 39% female). We designed a survey that comprehensively assessed multiple filtering criteria (1-7) across key assessment metrics (a-d).

Filtering criteria: As the most common filtering criterion in Information Systems (e.g., Barlow et al. 2018; Teubner and Flath 2019), we included a number of (1) attention checks: four items served as instructional manipulation checks ("You should not answer this item if you read it; it is to check your attention.": Paas et al. 2018, "While browsing online, have you ever had a lethal incident?": Paolacci et al. 2010; "please click on the Other as an option": Goodman et al. 2013; assessment of an image of a WhatsApp chat with a specific answer requested in the instructions: Peer et al. 2014) and four as factual manipulation checks (two each on reading an e-mail conversation, adapted from Downs et al. 2010, and two self-developed questions on a text on data protection in the European Union). Additionally, we measured participants' response time, both (2) for the whole questionnaire and (3) per page. We also measured reading speed (words per minute), again (4) for the whole questionnaire and (5) per page. Further, we also included two novel filtering criteria: first, we measured the (6) variance of the reading speed across pages, as measure of sudden inattention (e.g., browsing on another website). Second, we included (7) a standardized reading speed assessment at the beginning of the questionnaire. We used a standardized text offered by the iReST Study Group (Trauzettel-Klosinski and Dietz 2012). This text was openly positioned as reading speed assessment prior to the actual survey. Participants did not have to answer questions related to this text.

Assessment metrics: We assess the effect of the filters along four measures: (a) reliability is commonly assessed through Cronbach's alpha for multi-item scales and through test/retest correlation. As multi-item scales (which also served as filler items), we included six dimensions from the Information Privacy Concerns Scale (Malhotra et al. 2004), as well as measures capturing the intention to give personal information (Phelps et al. 2000) and consumer innovativeness (Parasuraman 2000). Of the Information Privacy measures, we used six items from the control dimension at the beginning and the end of the survey (e.g., "I believe other people are too much concerned with online privacy issues."), allowing for test/retest checks. For validity (b), we used the share of correct attention checks (both instructional and factual checks) as a measure. We then included two effects commonly used to assess (c) effect size: as within-subject measure, the anchoring effect, one of the most commonly replicated psychological effect (Tversky and Kahneman 1974, here, anchor: websites with user account; influenced: hours per day on the internet); and as between-subject measure, we adapted a priming task on mental accounting biases to rationality (Thaler 1985: different willingness-to-pay for the same product, when staying at a fancy hotel vs. cheap motel). Out of the observed effect size at different sample size levels, we computed (d) the post-hoc statistical power.

Structure: Participants started the survey with the standardized reading speed assessment. They then completed the first part of the test/retest items (page 0). The main body of the survey (page 1-5) contained the instructional manipulation and factual checks, multi-item scales as reliability measures and fillers and the effect size assessment. All components were fully randomized to prevent order effects. Page 6 contained the second part of the test/retest items as well as demographics.

Reading Speed and Attention across the Survey

An important debate in questionnaire design concerns the distribution of participant attention over time. For instance, some handbooks recommend placing key items for a research project early in the questionnaire (Aaker et al. 2013) while others favor a late positioning (Malhotra et al. 2012). Extant research finds that the average time per page decreases in the second half of a survey (Huang et al. 2012), potentially indicating a decline in attention. To better understand the relationship of answering time and attention, and its development over the survey, we first want to descriptively assess both.

Figure 1 Panels A and B map consumers reading speed (in words per minute: WPM) over the pages in the main body of the questionnaire (1-5). The density plots in Panel A show that while distribution is approximately normal on all pages, it is more dense at early and late stages of the questionnaire and shows higher variance in the middle (page 3). Panel B shows that reading speed and validity (share of correct attention checks) increase over the course of the survey. This could point to a correlation, but is likely also due to participants learning about the demands of the survey: after quickly completing the first page of the questionnaire (z-time = –1.58) and (relatively) often failing the attention check (z-validity = –1.55), participants adjust their reading speed as they notice that checks are included in the questionnaire and answering quality improves as of page 2. These results align with extant research: when participants become aware that attention checks are being used in a survey, they pay much closer attention to the questions (Hauser and Schwarz 2015; Oppenheimer et al. 2009), which explains the increase between the first and second page with attention checks. Both reading speed and validity subsequently improve. Please note that these differences are not due to the difficulty of the attention checks, as the latter's position was randomized. On an absolute level, however, both the reading speed (198-278 WPM) and correct attention checks (67 %-73%) are stable and align with previous findings of reading speed high answering quality on Amazon MTurk (Hauser and Schwarz 2016; Trauzettel-Klosinski and Dietz 2012). In summary, we find no indication that reading speed and attention strongly declines over a typical survey length (10 minutes, 7 content pages).

Panel C explores the relationship between question reading speed and validity. Participants are binned according to two reading speed measures (reading speed by question and the standardized reading speed test, iReST). We see that participants with an intermediate reading speed show the highest answer validity. Interestingly, and contrary to commonly held beliefs for time-based filters, fast participants are not necessarily worse than slow participants are. In fact, both distributions show a slight decline at slow reading speeds. This finding aligns with past research, which suggests that the professional survey takers on Amazon MTurk are highly attentive (Hauser and Schwarz 2016); reading speed, at least to a certain degree, can thus be a sign of proficiency.

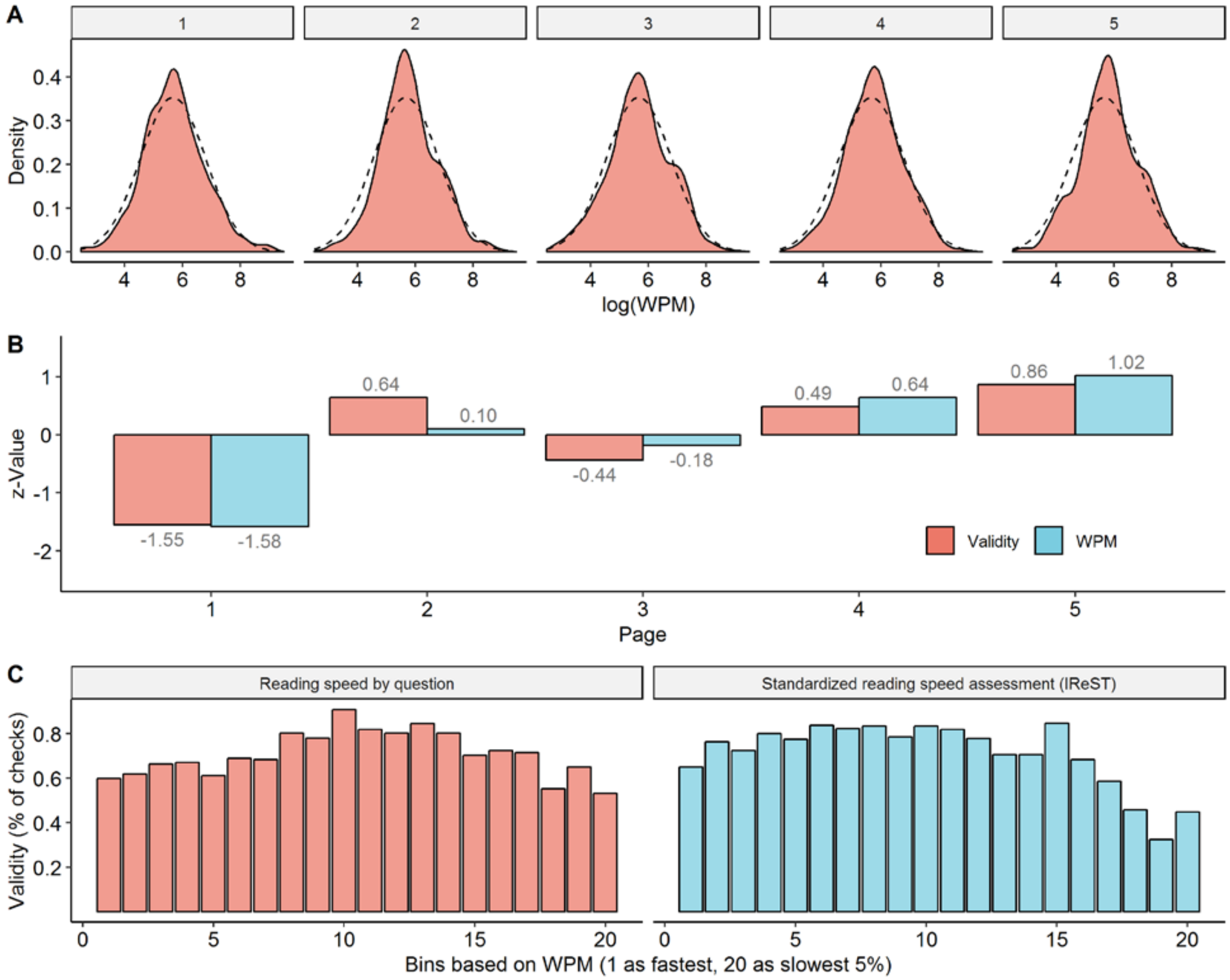

Figure 1. Reading speed and validity over survey page (A, B) and over different readings speed levels (C)

Comparison of Heuristic Filters

For each of the filters described above, we computed the metrics of interest (reliability, validity, effect size and power) while the discrimination threshold is varied. To enable a like-for-like comparison, we focused on the same thresholds across all filtering approaches: 10% of participants (20 of 200 excluded), 33% and 66%. Additionally, for total answering time we also include the common threshold of 2 standard deviations from the mean. Further, we compare all results against a baseline of randomly filtered participants. Table 1 provides a numeric overview of the filtering results at the thresholds, while Figure 2 provides an overview of the continuous development of the metrics (e.g., to identify drops or elbow points). Please note that we will discuss filtering by (8) machine learning and (9) logit models in the following section.

First, the most common filtering criteria – namely, (1) attention checks and (2) total time by questionnaire – perform well. (2) total time by questionnaire increases the validity of the results (0.60 at 10%), without reducing the other measures. However, effect sizes and power only remain at the level of random filters, suggesting that some parts of slower reading attention are directed towards spotting attention checks. Also filtering by (1) attention checks has a positive validity effect, but the increase in validity (.75 at 10%, .81 at 33%) is a necessary effect of filtering those participants that did not correctly answer the attention checks. We, therefore, do not use validity as a comparison criterion (see Table 1). Additionally, the effect size substantially increases for attention checks filters after an initial dip (.27 at 10%, .87 at 33%).

In contrast, filters by page ((3) total time and (5) reading speed) and (6) reading speed variance by page perform badly in terms of validity and reliability. Specifically, these filters reduce validity and reliability, although (6) reading speed variance has a positive effect on effect size and power (e.g., highest power of all filters at 10%: 0.84). This shows that filters that might reduce validity and reliability might appear attractive to researchers, because they bring out effects strongly. We, thus, would advise against the use of these filters.

Second, limited filtering (i.e., 10% of participants) does not substantially improve validity, reliability or effect size for most measures. In contrast, reliability falls below random filters for (3) time by page and (6) the reading speed variance by page. In addition, validity is low for these measures, although validity of (3) time by page at high filter levels increases again. Only (2) attention checks and (7) a standardized reading speed assessment before the questionnaire increase the validity at early filtering levels. There is no filtering effect on effect size and power at the 10% level, except for a slight increase for (7) standardized reading speed assessments. Hence, cautious filters, that only exclude a few participants, are often unlikely to improve the power of experimental results. Researchers with limited sample sizes who cannot lose too many respondents may want to revert to the (7) standardized reading speed test, which offers an early (but small) increase in validity, effect sizes and power.

Third, the selection of the filtering criterion has an effect on the estimated effect size. The estimated effect size varies substantially between small and large within the thresholds of common criteria (.30 to .94 for (3) Total Time by Page; .27 to .98 for (1) Attention Checks). Changes are particularly strong around elbow-points, which exist for some filtering criteria (e.g., attention checks: ~20% of participants; standardized reading speed: ~15% of participants). However, effect size differences arise also between criteria (at 33% level .31 for (3) Total Time by Page and .87 for (1) Attention Check filtering). This indicates that if the researcher has discretion over the selection of the filter, s/he might be strategic in their selection in order to obtain results that are more substantial.

In summary, we see that the most common filtering characteristics (total time and attention checks) perform equal to less common approaches. However, the most common current filtering heuristics only improve validity and reliability of the results at high filter levels. Also their effect on effect size and power varies substantially between filters and only arises when a substantial number of participants is filtered. This raises questions on whether the inattention or satisficing behavior, which is the reason for filtering, can reasonable be assumed to hold for half of the sample.

Extending Heuristic Filters: Machine Learning "Black Box" and Logit

The above results show that established filters leave the researcher with many degrees of freedom. Such "flexibility in data collection" (Simmons et al. 2011) may lead to post-hoc cleaning of the data to obtain the desired results (Chandler et al. 2014). Therefore, we test (8) a "black-box" approach using machine learning as a way to reduce researchers' degree of freedom over the experimental results. Specifically, the machine learning model would make use of all available filtering variables to predict metrics of interest (e.g., validity). As the research cannot decide which filtering criteria to use, but rather has to incorporate all potential variables in the machine learning dataset (e.g., as in Altmejd et al. 2019), s/he cannot be strategic in the variable and threshold selection. Even if the filtering is based on a specific model (e.g., (9) logit as in our study), the research still can influence the model parameters. Alternative approaches, such as a clustering of participants (Howell et al. 2017), also leave many decisions to the researcher (e.g., an interpretation of the segments). The black-box approach would not allow such adjustments. A similar approach has already been suggested – for much the same reasons – for assessing the replicability of experiments (Altmejd et al. 2019).

Further, the explanatory power of more complex, multivariate filters is likely to be higher. This might results in validity, reliability, effect size and power increases that exceed heuristic filters – which would be particularly interesting at low filtering levels (10%).

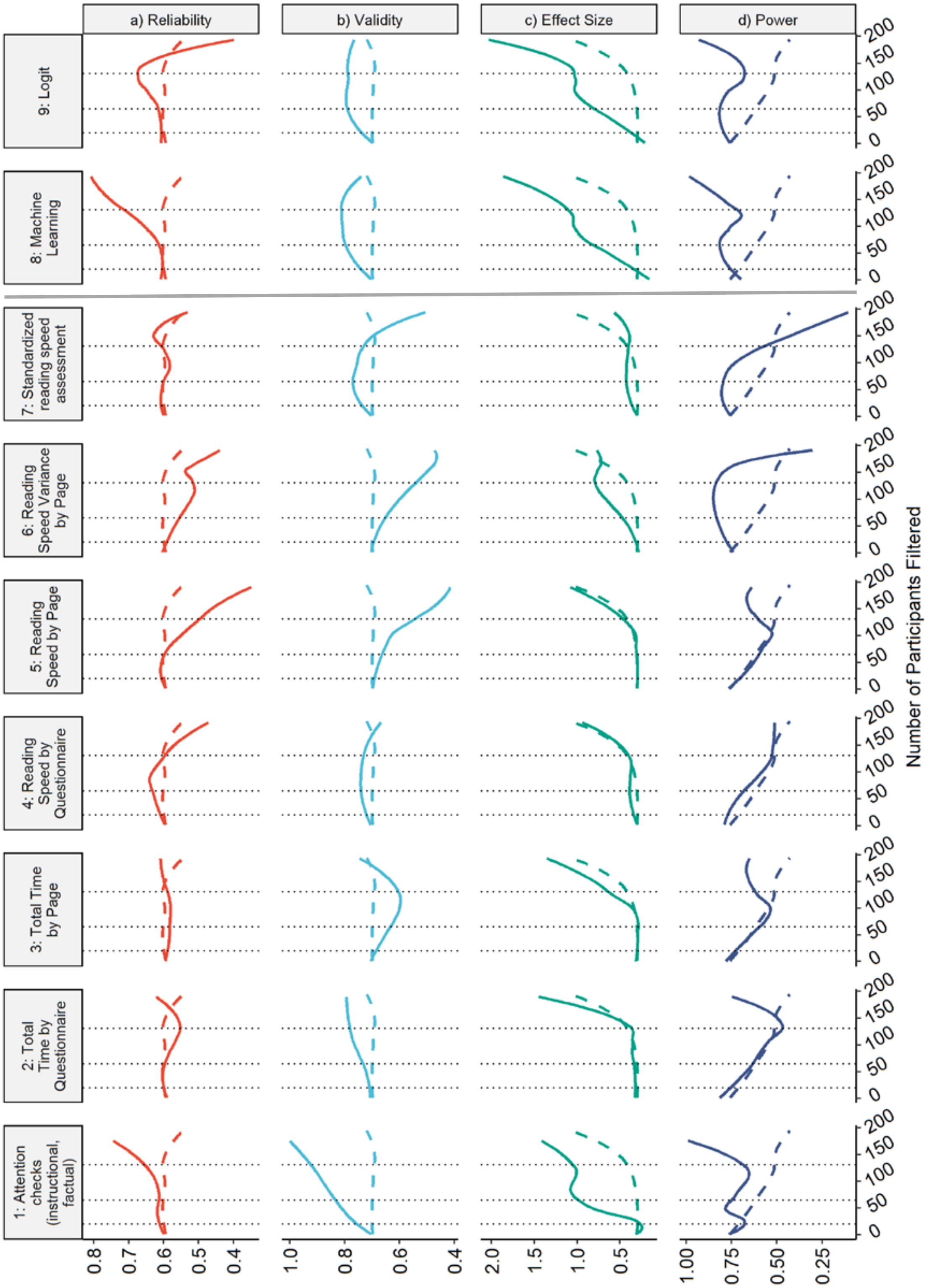

Figure 2. Effect of filtering criteria on key metrics (solid lines) vs. random filtering (dashed lines), thresholds (10%, 33%, 66%) marked

Filter	**Criterion**	**Threshold**	**(a) Reliability** (Cronbach's Alpha, Test/Retest Pearson's r)	**(b) Validity** (Share of IMCs)	**(c) Effect Size** (Cohen's d)	**(d) Power**
0	Random	10%	0.59	0.69	0.30	0.72
		33%	0.59	0.70	0.31	0.66
		66%	0.66	0.73	0.32	0.40
1	Attention checks (Instructional, factual)	10%*	**0.61**	0.75[a]	0.27	0.68
		33%*	**0.62**	0.81[a]	0.87	**0.81**
		66%*	0.65	0.91[a]	0.98	0.66
2	Total Time by Questionnaire	10%*	0.60	0.71	0.33	0.77
		33%	0.60	0.73	0.30	0.57
		66%*	0.55	0.77	0.36	0.46
		2sd (= 4 %)	0.60	0.71	0.33	0.79
3	Total Time by Page	10%*	0.59	0.69	0.30	0.73
		33%	0.58	0.64	0.31	0.59
		66%	0.57	0.61	0.94	0.73
4	Reading Speed by Questionnaire	10%	**0.61**	0.72	0.34	0.78
		33%	**0.62**	0.74	0.40	0.68
		66%	0.58	0.72	0.38	0.49
5	Reading Speed by Page	10%	**0.61**	0.69	0.29	0.68
		33%	0.60	0.66	0.31	0.60
		66%	0.50	0.56	0.40	0.53
6	Reading Speed Variance by Page	10%*	**0.61**	0.70	0.29	0.73
		33%	0.58	0.65	0.35	0.76
		66%	0.50	0.54	0.78	**0.82**
7	Standardized reading speed assessment	10%*	**0.61**	**0.74**	0.33	0.78
		33%*	0.61	0.77	0.42	0.80
		66%*	0.57	0.72	0.43	0.63
8	Machine Learning	10%	**0.61**	**0.74**	0.33	0.76
		33%	0.61	**0.80**	**0.93**	**0.81**
		66%	**0.70**	**0.81**	**1.07**	0.66
9	Logit	10%	**0.61**	**0.74**	**0.41**	**0.82**
		33%	0.61	0.79	**0.93**	**0.81**
		66%	0.67	0.78	1.06	0.65

Table 1. Effect of filtering criteria on key metrics at comparison thresholds: simple vs. multi-variate approaches; Note: Highest values at each threshold in bold; Values rounded down to the next observed values; [a]Direct consequence of filtering for IMCs – not used for comparison

Method: Our (8) “black box” machine learning approaches relies on a simple neural network, which predicts whether the response to a question will be correct or incorrect. We used question-level data from 18 features (including raw, logged and squared variables; see Figure 3) observed in the IMC questions for building and training a prediction model of the correct label (correct, incorrect). As multiple attention checks were included in the questionnaire, we obtained in total 1531 observations.

Data preprocessing included standardization of features and splitting the data into a training, validation and test set (20%/24%/56%). While the model will be trained on the training data, data from the validation set will be used to tune the hyper-parameters. Data from the test set will finally be used to evaluate the model performance. Further, as we are dealing with heavy class imbalance with regards to observed class membership in the data (i.e., ~70% of IMCs are correct), we used the Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al. 2002) to synthesize new instances in the minority class, effectively creating a balanced training set (i.e., test data that contains the same number of observations with correct and incorrect IMCs). This is especially important to create a conservative prediction model that performs equally well on classifying both, minority and majority class (He and Garcia 2009). The architecture of our model included an input layer, two hidden layers with 128 respective 64 neurons and an output layer. The design of the input layer corresponds to the number of features passed into the model for prediction, while the hidden layers are used to transform the relationships between features and labels, making them separable by a sigmoid function in the output layer (LeCun et al. 2015). Neurons in the layers were fully connected and activated only if their input exceeds 0 using the rectified linear unit (ReLU) function.

To prevent overfitting, we dropped neurons in each iteration and layer with a probability of 1/2, which is a configuration that is close to optimal for most networks (Srivastava et al. 2014). Since the features and labels fed into the model are generated by Amazon MTurk, which is a noisy process (Crump et al. 2013), we used the adaptive moment estimation algorithm (ADAM) to optimize the loss function in our model, which can effectively deal with noisy data and is computationally efficient (Kingma and Ba 2014). To increase reproducibility, we ran the model with default hyper-parameters (e.g., learning rate = .001). As loss function to be optimized by ADAM, we used binary-cross-entropy (i.e., log loss) as this metric penalizes misclassification in a binary target variable. We trained the model for 100 epochs (= number of times the test data is passed through the model) with a batch size of 15 (= number of observations passed into the model at once) training examples for quick convergence. The final model showed above average predictive performance (Accuracy: 75%; AUC = .81; F-Score: .82) on the test data.

After the model had been trained and evaluated on sub-samples of IMC questions, we used it to predict the quality of the remaining non-IMC questions. This process resulted in a conservative estimate of 1940 questions that were classified as “correct” and 2450 questions that were classified as “incorrect”. We used the probabilities of class membership to calculate an average score for each participant that reflects her propensity to give correct answers. Because the interpretation of machine learning results is difficult as coefficients cannot be directly interpreted as in traditional, linear modelling (Lipton 2016), we computed the absolute correlation of each variable included in the neural network with the predicted outcome (see Figure 3). Additionally, we complement the analysis through (9) a logistic regression model. Note that, although multiple interpretation methods exist (e.g., Model Reliance: Fisher et al. 2018; Shapley Value: Štrumbelj and Kononenko 2014) and produce insights into the black box, we relied on approximating the machine learning model with a “white box” logistic regression model. This model is interpretable (i.e., provides regression coefficients for each feature) and applicable as we did not expect complex non-linear effects in the underlying data. The model was fit on the same feature set and, consequently, enabled us to to draw implications about the underlying process (e.g., comparing feature importance) and to make findings more accessible.

Results: The machine learning algorithm ranks respondents in the order of highest predicted answering quality per respondent. Based on this ranking of participants, we can apply different filter thresholds (10%, 33% or 66%). As shown in Table 1 and Figure 2, especially the (8) machine learning model outperforms the simple filters. With regards to reliability, the machine learning is on par with other approaches at low filtering levels (10%: .61), but exceeds the latter at higher filtering levels (66%: .70). It is the only filter with increasing reliability over higher threshold levels. The machine learning model also shows the highest validity values at all thresholds (at least if we disregard the (1) attention checks, which filter by validity and, thus, necessarily have a stronger effect). Additionally, effect size and power is highest at intermediary and high filter levels. Additionally, most of the improvements are already achieved at filtering levels below 33%.

Filter based on (9) a logistic regression model perform similar to the (8) machine learning model, except for reliability. This positive result confirms our efforts of designing the logistic regression to approximate the machine learning model. Interestingly, both, machine learning and logit model show strong increases in effect size between 10% and 40% and beyond 66% of participants filtered, but somewhat of a plateau in-between.

This elbow, however, highlights a limitation of the current machine learning approach to filter participants: because a filtering threshold would also have to be selected (here: in terms of predicted answering quality), the researcher would again have open degrees of freedom. The present analysis only shows a way to make the filtering transparent. Ideally, the machine learning tool would also suggest an optimal cut-off threshold. Extant research has tested approaches to optimize hyper-parameter (for an overview see Bergstra et al. 2011), which could easily be transferred to our field of application. Our sample neural network, however, is trained only upon answering quality (i.e., IMC and factual manipulation checks), which offer no trade-off for an algorithm to consider (e.g., between validity and power).

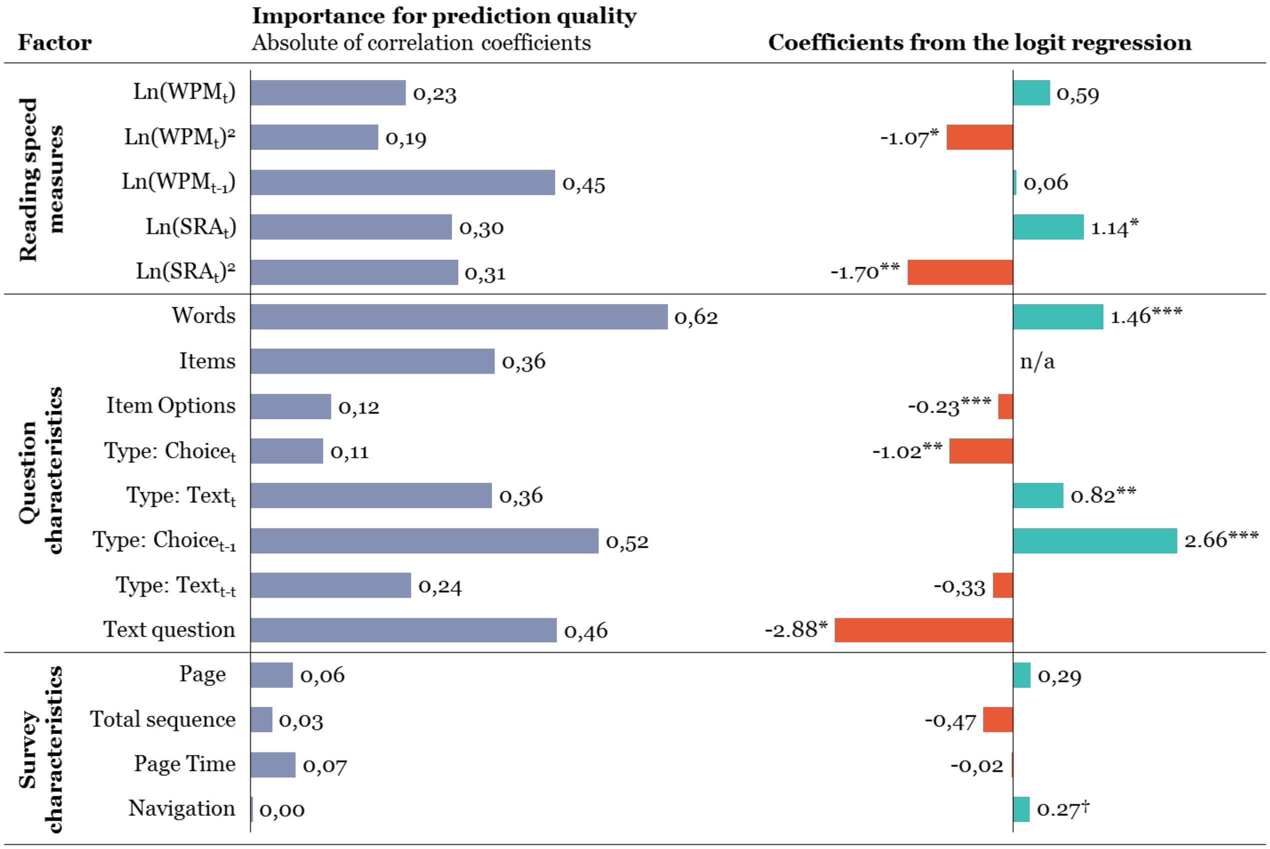

Figure 3. Estimated importance of different variables for attention-check prediction quality

Validity Drivers

As the machine learning model contains all potentially relevant variables, we can assess which variable is most important for predicting answering quality (i.e., validity). Because the machine learning "black box" is difficult to interpret, we compute an importance factor for each variable (absolute correlation coefficient between variable and predicted probability) and report the estimates of the logit model (see Figure 3). We refer to the logit model for easier interpretation.

The measures related to reading speed are most relevant for the discussion on answering time as filter and important for the predicted probabilities. Please note that reading speed is assessed by question (in contrast to the above filters on an overall or page level). Higher reading speed does not significantly influence the answering quality (b = .59, p = .16). Nevertheless, the quadratic reading speed effect is negative and significant ($b = -1.07$, $p < .05$). The same pattern also applies to the standardized reading assessment (SRA, linear: $b = 1.14$, $p < 0.05$, quadratic: $b = -1.70$, $p < 0.01$). This implies that the effect of reading speed follows an inverted U-shape: while the fast participants seem to be more attentive, participants with an intermediary reading speed have the highest answering quality. This result matches our previous descriptive observation (see Figure 1C).

As expected, question characteristics influence the answering quality: longer questions increase the odds of answering correctly ($b = 1.46$, $p < .001$), choice-based IMCs increase the probability of a correct answer ($b = 2.66$, $p < .001$), while text questions reduce the probability ($b = -2.88$, $p < .05$). Survey characteristics (e.g., page of the question) and overall characteristics (e.g., use of keyboard vs. mouse navigation) were not important for predicting the probability of correct answers, which is intuitive given the complete randomization of the questionnaire. Finally, the low importance and estimated effect of Page Time aligns with findings from the comparison of filters (see Table 1).

Discussion

Conclusion and Implications

A large share of current Information Systems research that uses Amazon MTurk, other online survey platforms and even screened pools of actual users of a service uses heuristic filters with inconsistent thresholds to identify inattentive participants. This may result in sub-optimal validity and reliability and a large variance in the observed effect, even with the same sample. To improve the validity and reliability of published results, as well as to ensure that we can identify actually existing effects, we need to compare different filtering approaches and consider alternative ways to filter potentially inattentive participants.

This research contributes to the challenge on three dimensions. First, we create transparency on the effect by comparing multiple heuristic filters (including novel ones, such as a reading speed test and reading speed variance) across a comprehensive set of relevant metrics for survey research (reliability, validity, effect size and power). We find that the most commonly used simple filters (overall answering time, attention checks) increase validity, but only at high filtering levels. However, results differ strongly in terms of effect size – both between filters and at different filtering thresholds. The more complex, multivariate filters outperform heuristic filtering. Especially the machine learning filter substantially increases validity and effect size, and at no loss of reliability or power. This suggests that multivariate, complex filters capture more aspects of respondent inattentiveness, while not falling prey to attention check spotting capabilities of professional survey takers.

The introduction of a machine learning "black box" model to assess participant attention is our second contribution. Because the model self-selects variables and their weight, researchers lose the discretion over the choice of the filter. This might prevent the selection of filters to maximize the attainable post-hoc effect size, potentially at the loss of validity (as in the reading speed variance filter example). Third, we contribute to the debate about optimal answering time in survey research. We find that fast completion time is not necessarily a sign of low attention. Rather, answering quality and answering speed follow an inverse U-shape, where the optimum lies at intermediate speed levels. As the overall quality declines with longer answering times, very fast participants seem to be more attentive than very slow ones.

These findings have implications for research practice. First, although the simplest filters produce valid and reliable results, a real improvement over random filtering in terms of effect size only arises at high filtering levels where less than a third of the paid sample is retained. As filtering by total response time requires no resources (i.e., in terms of additional question items), this filter is especially attractive for contexts where research efficiency is important (e.g., short studies, where attention checks would take up large part of the survey). Also, in line with extant research (Breitsohl and Steidelmüller 2018; Kung et al. 2018) we do not share concerns that attention checks reduce validity (e.g., from putting participants in a deliberative mindset), as they improve reliability and effect sizes in our data. Second, we would advise against the use of filters on a page level (e.g., reading speed by page), as these have low reliability and sometimes validity. Third, most of the traditional filters do not substantially increase validity, reliability and effect size at low levels of filtering. This might be an explanation why extant research often filters a high share of participants (e.g., 78% filtered: Barlow et al. 2018). We suggest, fourth, that research might use multivariate filtering approaches. Especially simple machine learning models are attractive, as they improve all variables of interest already at low filtering levels and because their “black box” character prevents strategic filtering (Altmejd et al. 2019). Journal editors and reviewers might use these results to decide upon the appropriateness of certain filtering procedures in submitted manuscripts.

Our findings also have specific implications for research using Amazon MTurk. First, as survey participants on the platform are often very professional, they are more sensitive to attention checks (Farrell et al. 2017; Hauser and Schwarz 2016) or share them in forums (Chandler et al. 2014; Wessling et al. 2017), diminishing the intended utility of attention checks. Our findings support this possibility: attention check answering quality was very high on average (70% correct) and increased over the course of the questionnaire – evidence for the presence of strategic behavior. If participants adjust their behavior after noticing that attention is being monitored and inattention might be a reason for rejected payment, we should assess participants’ quality by measures, which are not as easy to strategically adjust to (e.g., focus on spotting IMCs). For instance, filtering approaches are incentive aligned could be an alternative (e.g., by-page reading speed: participants have no incentive to wait on each page). However, all filters based on the page level perform worse than overall filters. Also relative changes in the answering behavior (e.g., (7) page variance) are not a valid alternative. In addition, the multivariate filters might be a solution, because they do not enable participants to use simple heuristics to game the filter. However, one might argue that the community of survey participants might also learn about these approaches, as about manipulation checks, and adjust its behavior in the long run (Hauser and Schwarz 2016). Further, here, adaptive “black box” neural networks might prevent an adjustment of the MTurk community.

These findings summarize to the following practical implications for researchers using online survey pools:

(1) MTurk participants are professional survey takers that have a high answering quality (in terms of attention checks)
(2) If a researcher wants to use a simple, widely accepted filtering heuristic, we recommend filters based on overall time and with commonly accepted thresholds (+/- 2 standard deviations) as these have high validity and reliability values
(3) If a researcher wants to signal that he/she did not use filtering for “p-hacking”, we recommend either (a) a pre-specification of the filtering approach (e.g., during a pre-registration of the study), or (b) the use of a “black box” machine learning approach to filter based on answering quality

If researchers want to go beyond retrospective filtering, a potential application of our approach pertains to the automatic in-survey assessment participant attention in order to deter inattention in the first place. Extant research suggests that confronting participants with failed attention checks subsequently increases attention (Hauser and Schwarz 2016; Oppenheimer et al. 2009). Using current reading speed, or even a page-wise assessment from a machine learning algorithm could be an option to continuously monitor behavior in a less apparent way (as, e.g., for reviews: Kumar et al. 2018): only if participants progress too quickly or show other learned signs of inattention, they could be automatically warned, for instance through a pop-up. This might also reduce the necessity of costly post-hoc filtering of participants, which have already completed a survey. As pre-trained model can easily be deployed to web applications (e.g., tesorflow.js), implementing this solution in the field of online surveys would be a straightforward task.

Limitations and Future Research

A large share of research studies are currently conducted on professional survey platforms, such as Amazon Mturk – our study as well. This poses a specific limitation to the findings. Even when acknowledging the general representativeness and validity of MTurk findings (Cheung et al. 2017), research notes that MTurkers are often more strategic and professional in their answering. Our findings support this assertion, in that participants with the highest share of correct answers were also the fastest participants. This is likely a consequence of the professionalism in the participation in surveys. It, thus, would be highly interesting if the same factors for answering quality emerge when applying our neural network to other – potentially less experienced – groups of participants, such as students.

Further, the complexity of our modelling could be improved. Multiple roads are possible here, which we plan to address in the near future. First, we train the machine learning model on answering quality on a question level, but a multiplicative measure which takes into account all measures of interest (i.e., also reliability, effect size, power) would better capture the tradeoffs associated with filtering (e.g., between effect size and power). Reliability, effect size and power, however, can only be computed on a subject (and not a question level), substantially reducing the sample size for the model. Subsequent research could train a model on a multiplicative measure, but would require a larger sample. Second, our current approach investigates simple, feed-forward machine learning applications (Altmejd et al. 2019; Kumar et al. 2018) that treats the training examples as independent (except for lagged variables), without controlling for the repeated nature of data generation. Consequently, the time-dependent global answer pattern of a subject (e.g., random intercept in a generalized linear mixed model) might contribute to increasing the model performance. Using the architectural capabilities of, for example, Recurrent Neural Networks and specifically Long Short-Term Memory Units (LSTMs) future research could aim at improving the filter quality even further. As a caveat, however, such modeling would substantially increase the analytical complexity for the researcher, making the approach less applicable in research practice. Third, we make the suggestion that a "black box" machine learning approach might be beneficial because it reduces researchers' degrees of freedom.

Despite selecting variables and weights, however, the approach in its current form is still contingent on a threshold for the filtering. This again may lead to strategic behavior. Future models, therefore, should also incorporate a selection of the optimal filtering threshold (e.g., as in the choice for optimal hyper-parameter; potentially trading off validity and effect size). Potentially, to eliminate researcher discretion, scientific instances could aim for installing a collectively designed model of survey quality to go along with publication requirements. Finally, survey participation is strongly influenced by the participants' setting (Paas et al. 2018). As many contextual variables are available (e.g., device, screen size, location of the IP address), the machine learning model might be enriched to further improve prediction quality.

Points for Discussion

We hope that this research helps to encourage a discussion about the present methodological paradigm of filtering participants for survey research in Information Systems and beyond. Specifically, our findings raise four, more fundamental, questions that our field should address:

- Can there be such a thing as a "silver bullet" filter? Alternatively, do filters differ by context (e.g., required resource efficiency)?
- Which measures should a manuscript report when filtering (e.g., reduction in power, changes in validity, reliability, etc.)? Moreover, would additional rigor in reporting lead to less relevant results?
- More ontologically, do we want to continue to treat the researcher as academic brute, whose freedom needs to be curtailed by some form of a research contract?
- Finally, is it appropriate to use "black box" machine learning models to curb researcher degrees of freedom through selecting the filters as well as the thresholds? Especially, do we want to relinquish control of the rules of what is methodological appropriate?

References

Aaker, D. A., v. Kumar, Leone, R. P., and Day, G. S. 2013. *Marketing Research*, Hoboken, NJ: Wiley.

Ahn, J.-H., Bae, Y.-S., Ju, J., and Oh, W. 2018. "Attention Adjustment, Renewal, and Equilibrium Seeking in Online Search: An Eye-Tracking Approach," *Journal of Management Information Systems* (35:4), pp. 1218–1250.

Altmejd, A., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., and Camerer, C. 2019. "Predicting the Replicability of Social Science Lab Experiments," *MetaArXiv*.

Barber, L. K., Barnes, C. M., and Carlson, K. D. 2013. "Random and Systematic Error Effects of Insomnia on Survey Behavior," *Organizational Research Methods* (16:4), pp. 616–649.

Barlow, J. B., Warkentin, M., Ormond, D., and Dennis, A. R. 2018. "Don't Even Think About It! The Effects of Antineutralization, Informational, and Normative Communication on Information Security Compliance," *Journal of the Association for Information Systems* (19:8), pp. 689–715.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Balázs Kégl 2011. "Algorithms for Hyper-Parameter Optimization," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger (eds.): Curran Associates, Inc, pp. 2546–2554.

Breitsohl, H., and Steidelmüller, C. 2018. "The Impact of Insufficient Effort Responding Detection Methods on Substantive Responses: Results from an Experiment Testing Parameter Invariance," *Applied Psychology* (67:2), pp. 284–308.

Bui, D., Myerson, J., and Hale, S. 2015. "Age-Related Slowing in Online Samples," *Psychological Record* (65:4), pp. 649–655.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience* (14:5), pp. 365–376.

Chandler, J., Mueller, P., and Paolacci, G. 2014. "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers," *Behavior Research Methods* (46:1), pp. 112–130.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research* (16), pp. 321–357.

Cheung, J. H., Burns, D. K., Sinclair, R. R., and Sliter, M. 2017. "Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations," *Journal of Business and Psychology* (32:4), pp. 347–361.

Clemons, E. K., Wilson, J., Matt, C., Hess, T., Ren, F., Jin, F., and Koh, N. S. 2016. "Global Differences in Online Shopping Behavior: Understanding Factors Leading to Trust," *Journal of Management Information Systems* (33:4), pp. 1117–1148.

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. 2013. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research," *PloS one* (8:3), pp. 1-18.

Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. 2010. "Are your Participants Gaming the System?" *Proceedings of the 28th Annual CHI Conference on Human Factors in Computing Systems*, pp. 2399–2402.

Farrell, A. M., Grenier, J. H., and Leiby, J. 2017. "Scoundrels or Stars? Theory and Evidence on the Quality of Workers in Online Labor Markets," *The Accounting Review* (92:1), pp. 93–114.

Fisher, A., Rudin, C., and Dominici, F. 2018. "All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance," *arXiv preprint*.

Goodman, J. K., Cryder, C. E., and Cheema, A. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making* (26:3), pp. 213–224.

Goodman, J. K., and Paolacci, G. 2017. "Crowdsourcing Consumer Research," *Journal of Consumer Research* (44:1), pp. 196–210.

Hauser, D. J., and Schwarz, N. 2015. "It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks," *SAGE Open*.

Hauser, D. J., and Schwarz, N. 2016. "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than do Subject Pool Participants," *Behavior Research Methods* (48:1), pp. 400–407.

He, H., and Garcia, E. A. 2009. "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering* (21:9), pp. 1263–1284.

Howell, J. R., Ebbes, P., Liechty, J., and Jenkins, P. 2017. "Gremlins in the Data: Identifying the Information Content of Research Subjects," *HEC Paris Research Paper*, pp. 1–51.

Huang, J. L. 2014. "Does Cleanliness Influence Moral Judgments? Response Effort Moderates the Effect of Cleanliness Priming on Moral Judgments," *Frontiers in Psychology* (5), pp. 1–8.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., and DeShon, R. P. 2012. "Detecting and Deterring Insufficient Effort Responding to Surveys," *Journal of Business and Psychology* (27:1), pp. 99–114.

Kershner, A. M. 1964. "Speed of Reading in an Adult Population under Differential Conditions," *Journal of Applied Psychology* (48:1), pp. 25–28.

Kingma, D. P., and Ba, J. 2014. "Adam: A Method for Stochastic Optimization," *arXiv preprint*.

Krosnick, J. A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology* (5:3), pp. 213–236.

Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2018. "Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning," *Journal of Management Information Systems* (35:1), pp. 350–380.

Kung, F. Y.H., Kwok, N., and Brown, D. J. 2018. "Are Attention Check Questions a Threat to Scale Validity?" *Applied Psychology* (67:2), pp. 264–283.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep learning," *Nature* (521:7553), pp. 436–444.

Lipton, Z. C. 2016. “The Mythos of Model Interpretability,” *arXiv preprint*.

Malhotra, N. 2008. “Completion Time and Response Order Effects in Web Surveys,” *Public Opinion Quarterly* (72:5), pp. 914–934.

Malhotra, N. K., Birks, D. F., and Wills, P. 2012. *Marketing Research*: *An Applied Approach*, Harlow: Financial Times Prentice Hall.

Malhotra, N. K., Kim, S. S., and Agarwal, J. 2004. “Internet Users’ Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model,” *Information Systems Research* (15:4), pp. 336–355.

Mason, W., and Suri, S. 2012. “Conducting Behavioral Research on Amazon’s Mechanical Turk,” *Behavior Research Methods* (44:1), pp. 1–23.

Moore, S. G. 2015. “Attitude Predictability and Helpfulness in Online Reviews: The Role of Explained Actions and Reactions,” *Journal of Consumer Research* (42:1), pp. 30–44.

Open Science Collaboration 2015. “Estimating the Reproducibility of Psychological Science,” *Science* (349:6251), pp. 943–951.

Oppenheimer, D. M., Meyvis, T., and Davidenko, N. 2009. “Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power,” *Journal of Experimental Social Psychology* (45:4), pp. 867–872.

Owens, J., and Hawkins, E. M. 2019. “Using Online Labor Market Participants for Nonprofessional Investor Research: A Comparison of MTurk and Qualtrics Samples,” *Journal of Information Systems* (33:1), pp. 113–128.

Paas, L. J., Dolnicar, S., and Karlsson, L. 2018. “Instructional Manipulation Checks: A longitudinal analysis with implications for MTurk,” *International Journal of Research in Marketing* (35:2), pp. 258–269.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. 2010. “Running Experiments on Amazon Mechanical Turk,” *Judgment and Decision Making* (5), pp. 411–419.

Parasuraman, A. 2000. “Technology Readiness Index (Tri),” *Journal of Service Research* (2:4), pp. 307–320.

Peer, E., Vosgerau, J., and Acquisti, A. 2014. “Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk,” *Behavior Research Methods* (46:4), pp. 1023–1031.

Phelps, J., Nowak, G., and Ferrell, E. 2000. “Privacy Concerns and Consumer Willingness to Provide Personal Information,” *Journal of Public Policy & Marketing* (19:1), pp. 27–41.

Schmidt, G. B. 2015. “Fifty Days an MTurk Worker: The Social and Motivational Context for Amazon Mechanical Turk Workers,” *Industrial and Organizational Psychology* (8:02), pp. 165–171.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant,” *Psychological Science* (22:11), pp. 1359–1366.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. 2014. “P-Curve: A Key to the File-Drawer,” *Journal of Experimental Psychology: General* (143:2), pp. 534–547.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 2014. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research* (15), pp. 1929–1958.

Stokel-Walker, C. 2018. *Bots on Amazon’s Mechanical Turk are ruining psychology studies*. https://www.newscientist.com/article/2176436-bots-on-amazons-mechanical-turk-are-ruining-psychology-studies. Accessed 5 April 2019.

Štrumbelj, E., and Kononenko, I. 2014. “Explaining Prediction Models and Individual Predictions with Feature Contributions,” *Knowledge and Information Systems* (41:3), pp. 647–665.

Teubner, T., and Flath, C. M. 2019. “Privacy in the Sharing Economy,” *Journal of the Association for Information Systems* (20:3), pp. 213–242.

Thaler, R. 1985. “Mental Accounting and Consumer Choice,” *Marketing Science* (4:3), pp. 199–214.

Trauzettel-Klosinski, S., and Dietz, K. 2012. “Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST,” *Investigative Ophthalmology & Visual Science* (53:9), pp. 5452–5461.

Tversky, A., and Kahneman, D. 1974. “Judgment under Uncertainty: Heuristics and Biases,” *Science* (185:4157), pp. 1124–1131.

van Herk, H., Poortinga, Y. H., and Verhallen, T. M. M. 2004. “Response Styles in Rating Scales:Evidence of Method Bias in Data From Six EU Countries,” *Journal of Cross-Cultural Psychology* (35:3), pp. 346–360.

Warkentin, M., Goel, S., and Menard, P. 2017. “Shared Benefits and Information Privacy: What Determines Smart Meter Technology Adoption?” *Journal of the Association for Information Systems* (18:11), pp. 758–786.

Wessling, K. S., Huber, J., and Netzer, O. 2017. “MTurk Character Misrepresentation: Assessment and Solutions,” *Journal of Consumer Research* (44:1), pp. 211–230.

Zhang, C., and Conrad, F. 2014. “Speeding in Web Surveys: The Tendency to Answer Very Fast and its Association with Straightlining,” *Survey Research Methods* (8:2), pp. 127–135.