APPLIED PSYCHOLOGY

THE INTERNATIONAL ASSOCIATION OF APPLIED PSYCHOLOGY

APPLIED PSYCHOLOGY: AN INTERNATIONAL REVIEW, 2018, *67* (2), 264–283
doi: 10.1111/apps.12108

# Are Attention Check Questions a Threat to Scale Validity?

Franki Y.H. Kung*, Navio Kwok and Douglas J. Brown

*University of Waterloo, Canada*

Attention checks have become increasingly popular in survey research as a means to filter out careless respondents. Despite their widespread use, little research has empirically tested the impact of attention checks on scale validity. In fact, because attention checks can induce a more deliberative mindset in survey respondents, they may change the way respondents answer survey questions, posing a threat to scale validity. In two studies, we tested this hypothesis ($N = 816$). We examined whether common attention checks—instructed-response items (Study 1) and an instructional manipulation check (Study 2)—impact responses to a well-validated management scale. Results showed no evidence that they affect scale validity, both in reported scale means and tests of measurement invariance. These findings allow researchers to justify the use of attention checks without compromising scale validity and encourage future research to examine other survey characteristic-respondent dynamics to advance our use of survey methods.

## INTRODUCTION

Self-report measurement scales are the capstone method in survey research and they influence many organisational decisions, such as personnel selection and assessment. In an ideal world, survey respondents are assumed to pay adequate attention to each scale item so that their responses are meaningful and offer valid measurement of a psychological construct. Yet, the ideal world differs from the reality. Some evidence suggests that at least 5 per cent or more of respondents answer scale items carelessly (Johnson, 2005), and this percentage can be as high as 60 per cent when respondents receive little or no incentive to complete a survey (Berry et al., 1992; Hauser & Schwarz, 2016; Meade

* Address for correspondence: Franki Kung, Department of Psychology, University of Waterloo, 200 University Ave. W., Ontario, Canada N2L3G1. Email: franki.kung@uwaterloo.ca

This research was supported by Social Sciences and Humanities Research Council of Canada through the Vanier Canada Graduate Scholarship (CGV-SSHRC-00379) (to FK), Joseph-Armand Bombardier Canada Graduate Doctoral Scholarship (SSHRC CGS-D 767-2016-1247) (to NK) and the Canadian SSHRC Grant (to DB).

& Craig, 2012). These careless responses directly challenge the validity of any scale measurement, and as a result, can lead to misleading findings and conclusions (e.g. Bowling et al., 2016; Huang, Liu, & Bowling, 2015; Maniaci & Rogge, 2014).

To ensure scale validity, many researchers recommend including attention check questions in surveys (e.g. Berinsky, Margolis, & Sances, 2014; Curran, 2016; Huang, Liu, et al., 2015). These attention checks are items usually embedded early in a survey with an obvious correct response. Their purpose is to identify careless respondents and allow researchers to screen them out prior to conducting analyses (Maniaci & Rogge, 2014; Schmitt & Stults, 1985). For example, on a scale from 1 to 5, an item that reads, "please select four for this item", assesses whether a respondent actually pays attention when reading the item. Because they are a low-cost and efficient method to protect scale validity, attention checks are now widely employed and considered to be a desirable feature in survey designs across disciplines (Berinsky et al., 2014; Bowling et al., 2016; Hauser & Schwarz, 2016).

Despite their benefits, attention checks have limitations (e.g. Curran, 2016; Curran & Hauser, 2015). In fact, very recent findings showed that the inclusion of attention checks caused respondents to approach subsequent questions differently (Hauser & Schwarz, 2015). This raises the concern of a novel systematic threat to scale validity that the past literature has not explored, which is the focus of our current research—verifying whether the threat is real. This investigation is critical not only because it advances our understanding of the use of attention checks and survey methods in general, but also because it provides direct evidence to support (or oppose) the use of attention checks in the future. In the following sections, we will describe two popular types of attention checks, discuss why they may affect scale validity, and then present findings from two studies each examining the effect of one type of attention check on the scale validity of a representative and influential management scale.

## Two Popular Attention Checks

Researchers use of attention checks may vary slightly, but they converge on two major forms. Perhaps the most popular form of attention check is instructed-response items (see Bowling et al., 2016; Meade & Craig, 2012; Ward & Pond, 2015), which are items embedded in a scale with an obvious correct answer. The earlier example—"please select four for this item" and "please select moderately inaccurate for this item" (Huang, Curran, Keeney, Poposki, & DeShon, 2012)—are common instructed-response items. Because anyone who has read the item should be able to answer the item correctly, wrong answers to the instructed-response item indicate inattention. This assumption is not always perfect as some may suggest other possible explanations for a wrong answer (Curran & Hauser, 2015). However, instructed-response items,

in general, have shown great success in screening out careless survey respondents to protect the validity of scale measurement (see Meade & Craig, 2012; Woods, 2006).

Another variation of attention checks is called instructional manipulation checks (IMC; Oppenheimer, Meyvis, & Davidenko, 2009). An IMC item tends to be elaborate, with a critical cue to the correct answer hidden within a lengthy instruction. Appendix A provides a typical example of an IMC item. In that example, instead of answering the surface question (i.e. what workplace facilities are available), the key to passing the IMC is the last sentence of the paragraph, which instructs respondents to enter "I read the instructions" in the textbox. The assumption of why IMC traps careless respondents is similar to that of instructed-response items. In this respect, anyone who has read the entire set of instructions should be able to follow the "real" instruction in the last sentence, while any other response indicates inattention. Compared to instructed-response items, an IMC requires more effort in reading and hence, in theory, is a more effective technique for identifying careless respondents. However, because an IMC stands out from the typical survey questions, its format limits the utility of reusing an IMC, particularly in the same sample. Moreover, an IMC looks more elaborate and can seem trickier to participants, which may influence a respondent's subjective survey experience more strongly.

Since its publication in 2009, the IMC has been cited over 880 times.[1] Despite the increasingly popular use of attention checks, there has been a paucity of research on how they influence survey responses. In particular, attention checks may have a systematic influence on they way respondents answer and understand the actual survey questions (Hauser & Schwarz, 2015).

## Why Attention Checks Could Be a Threat to Scale Validity

Attention checks may alter scale responses, influencing scale validity, due to their binding nature of explicitness. To ensure that a wrong answer reflects mostly—if not solely—inattention, the correct response to an attention check has to be very explicit (Curran, 2016). With an instructed-response item that states, "please select four for this item", the correct answer of "4" is clear, and any other alternatives imply respondents inattention. Because of this required explicitness, survey respondents who read the attention check can tell it is a "trap" in the survey.

Survey respondents resemble lay scientists and actively infer researchers intentions from the survey questions (see Schwarz, 1994, 1999). When respondents see attention checks, they infer that researchers want to know

[1] As of March 2017.

whether they are paying attention—a goal that may otherwise not be activated in their mind (Hauser & Schwarz, 2015). Therefore, an attention check question in and of itself can trigger respondents to be particularly deliberate when filling out the remainder of a survey. Ample research has already shown that when people are in a more deliberative mindset, they show very different judgment and decision-making processes (e.g. Frederick, 2005; Stanovich & West, 2000, 2008). Attention checks may, therefore, alter people's response behaviors and cause survey respondents to provide inconsistent or inaccurate answers that deviate from their typical responses. In other words, attention checks can introduce response variance in a scale that is not part of the target construct, threatening the validity of the scale.

One theory that explains why deliberation can alter survey responses comes from research demonstrating that conscious deliberation, compared to intuition, causes overthinking—more biased and inconsistent information-processing (Tordesillas & Chaiken, 1999; Wilson & Schooler, 1991). Some also refer to this phenomenon as the deliberation-without-attention effect (Dijksterhuis, Bos, Nordgren, & van Baaren, 2006)—the idea that people make worse decisions when they deliberate compared to when they simply follow their "gut feeling". Although there has been a debate about the exact nature of the effect, such as whether deliberation can happen unconsciously or not (e.g. Custers, 2014; Mamede et al., 2010; Nordgren & Dijksterhuis, 2009), the basic premise here is that proactive deliberation *can* induce inferior judgment. The process of deliberation invites more thoughtful recollection and consideration of information; nevertheless, in reality, available information for decision-making is limited, and often irrelevant. Because of the bias that people inflate the importance of information that is available at the moment, the more people deliberate, the more they suffer from suboptimal weighting of importance of information (Levine, Halberstadt, & Goldstone, 1996; Wilson, Hodges, & LaFleur, 1995). As a result, deliberation can lead to worse judgment, often producing either inaccurate or inconsistent decisions (e.g. Dijksterhuis et al., 2006; Nordgren & Dijksterhuis, 2009; Tordesillas & Chaiken, 1999; cf. Calvillo & Penaloza, 2009).

Generally speaking, the effect of deliberation on decisional inaccuracy and inconsistency is evident especially when a decision is complex and multi-faceted (Dijksterhuis et al., 2006). However, this contingency does not preclude the fact that deliberation may have a nontrivial impact on seemingly simple decision-making, like survey responses. In fact, consumer behavior research has recognised the pervasive effect of deliberation on day-to-day choices as simple as people s preferences of pictures, jams, and jelly-beans (e.g. Nordgren & Dijksterhuis, 2009). In the context of survey research, perhaps that is also why instructions for a number of measurement scales encourage the respondents to respond intuitively, to not worry about answers being right or wrong, and to respond with the first thought that

comes to mind (e.g. Ferguson, Matthews, & Cox, 1999; Van Lange, 1999).[2] Yet, whether attention checks in fact lead participants to deliberate more and overthink, resulting in more inaccurate or inconsistent scale ratings, remains an empirical question to explore.

Perhaps the most direct evidence—that supports the idea that attention checks can pose a threat to scale validity—are studies showing that attention checks indeed cause respondents to deliberate more. Hauser and Schwarz (2015) conducted studies comparing the effect of receiving (vs. not receiving) an IMC on respondents subsequent performance on a deliberation task. Their results indicated that those who had been given an IMC before the deliberation task scored higher on deliberation; for instance, they spent more time thinking about a solution, and relying less on intuitive and more on rational reasoning. Moreover, these differences did not depend on how familiar the respondents were with attention checks, which means that seeing the attention check for the first time has the same effect as having seen the attention check previously. Taken together, these initial findings raise the possibility that attention checks may damage scale validity.

Ironically, if the concern that attention checks threaten scale validity is real, it does not influence the survey experience of careless respondents (i.e. those who do not notice the attention check); rather, it is those who are careful and whose data are likely retained in the actual analysis who will be affected. This outcome can be disastrous because it means that the recommended practice of using attention checks for screening (Buhrmester, Kwang, & Gosling, 2011; Paolacci & Chandler, 2014) may create a more serious problem than it solves. Typically, careless participants are not the majority of a sample and error due to careless responding tends to be random (e.g. Johnson, 2005). Because it is random, error variance due to careless responding can possibly be attenuated when empirical evidence accumulates and the overall sample gets larger. However, the error due to attention checks, if true, can be more *systematic*. It means that the error accumulates across studies and cannot be attenuated even by having a larger sample size. Such systematic error is a confound that can sway results in a particular direction and bias research conclusions. While using attention checks to screen participants is getting increasingly popular as a "best practice" in the field (DeSimone, Harms, & DeSimone, 2015), their potential threat to scale validity is notably also getting increasingly critical to address.

Are attention checks a threat to scale validity? This is an empirical question that remains open and requires more evidence to address. Our current research answers this question and directly examines whether the inclusion

[2] We thank a reviewer for this observation.

of attention checks influences scale responses. There are two conventional ways to compare scale responses: First, whether respondents differ in their answers to the scale, which would be straightforward, looking at any systematic mean differences in scale scores (e.g. *t*-test). Second, whether respondents differ in how they construe the meaning of the items of the scale, which can be examined with measurement invariance tests using structural equation modeling (e.g. Vandenberg & Lance, 2000; Woods, 2006). All else equal, if it is true that attention checks are a threat to scale validity, we should observe that a group of survey respondents who have received an attention check (vs. a group that does not) prior to answering a scale differs in their scale score, and/or the scale fails invariance tests across the two respondent groups. However, if attention checks are not a threat, we should observe no significant difference in the scale score, and the scale should achieve measurement invariance across the two respondent groups. These comparisons motivate the design of our studies as described in more detail below, and their results will provide direct initial evidence to address whether attention checks are a potential threat to scale validity.

## STUDY 1

This study tests whether embedding instructed-response items in a scale or not influences people's responses to that scale. For the purpose of this study, we utilised a popular organisational citizenship behavior (OCB) scale developed by Podsakoff, MacKenzie, Moorman, and Fetter (1990). Two criteria guided our choice of this scale. First, with a goal to inform the management literature, we wanted to examine the impact of attention checks on an influential scale. This OCB scale developed by Podsakoff and colleagues fits this criterion as it is highly cited and widely used.[3] A second reason for choosing this scale is that it is multidimensional. Compared to unidimensional scales, scales with multiple sub-dimensions are more nuanced and should be more sensitive to varying scale responses across groups of respondents. Therefore, to aim for a stronger test of any potential influence of attention checks, we searched for a multidimensional scale. Based on the criteria, we selected the organisational citizenship behavior scale.

### Method

*Participants and Procedures.* We recruited participants through Amazon's Mechanical Turk© (MTurk; Buhrmester et al., 2011) to complete a 5-minute online survey about workplace experiences for US$0.30. We set the location restriction to only be in the US and specifically preselected

[3] Cited over 4,800 times as of March 2017.

English-as-first-language full-time workers. These restrictions ensure that the scales in the survey are relevant to the respondents and minimise differences in first language and nationality that might otherwise contaminate scale interpretation. Survey respondents first reported their demographics and were then randomly assigned to either the experimental condition with instructed-response items embedded in the OCB scale or the control condition that had no instructed-response items. To ensure enough power for measurement invariance tests, we aimed to recruit at least 300 participants in total and collected a final sample of 451 participants. Detailed demographics are reported in Table 1.[4]

*Organisational Citizenship Behavior.* Respondents answered questions in the 24-item OCB scale (Podsakoff et al., 1990) (e.g. "Helps others who have been absent" from 1 = *Strongly Disagree* to 7 = *Strongly Agree*) ($\alpha = .91$). The scale has five dimensions, including conscientiousness, sportsmanship, virtue, courtesy, and altruism. Each dimension has four to five items ($\alpha$s ranged from .75 to .82).

*Instructed-Response Items.* Respondents in the experimental condition answered two additional instructed-response items embedded in the OCB scale (i.e. "For this question, please select number two to demonstrate your attention" and "For this question, please select number six to demonstrate your attention"; see exact materials in Appendix B).

## Results

*Mean Differences.* To examine whether the instructed-response items influenced scale responses, we first compared the mean scores of the scale across the two conditions. Results of a between-subjects ANOVA indicated that the experimental and control conditions did not differ in overall mean scores or mean scores within each sub-dimensions, $ps > .08$ (see Table 1). These results suggest that attention checks did not significantly alter the degree to which respondents endorsed the items.

*Measurement Invariance.* To assess whether the instructed-response items influenced the way respondents construe the meaning of the scale items, we conducted measurement invariance tests using structural equation

[4] In the experimental condition, 21 participants answered either one or both instructed-response items incorrectly. These participants reported significantly lower means than the participants who successfully completed the attention check and in the control condition for the overall OCB scale and most sub-dimensions. Including these participants or not did not change the overall patterns of results. For a more stringent test of the hypotheses, they were excluded from the analysis.

TABLE 1
Summary of Demographics, Conditional Means, and Standard Deviations

| | *Study 1* | | | *Study 2* | | |
|---|---|---|---|---|---|---|
| | *Experimental* | *Control* | *d* | *Experimental* | *Control* | *d* |
| Mean scores | | | | | | |
| OCB overall | 5.64 (.67) | 5.54 (.71) | .14 | 5.33 (.79) | 5.29 (.85) | .05 |
| Conscientiousness | 5.71 (1.01) | 5.70 (.93) | .01 | 5.61 (.97) | 5.57 (1.05) | .04 |
| Sportsmanship | 5.45 (1.05) | 5.39 (1.13) | .06 | 4.72 (.87) | 4.72 (.85) | .00 |
| Virtue | 5.50 (1.11) | 5.39 (1.10) | .10 | 5.26 (1.20) | 5.05 (1.31) | .17 |
| Courtesy | 6.00 (.80) | 5.87 (.89) | .15 | 5.59 (.97) | 5.59 (1.02) | .00 |
| Altruism | 5.46 (.58) | 5.37 (.62) | .15 | 5.43 (1.09) | 5.46 (1.13) | .03 |
| Mean age | 32.50 (8.99) | 33.38 (9.18) | | 34.09 (10.88) | 34.16 (10.89) | |
| Male % | 43.6 | 40.9 | | 52.9 | 51.8 | |
| Mean tenure (in years) | 4.61 (5.26) | 5.17 (5.69) | | 4.89 (4.18) | 5.73 (5.45) | |
| Race (%) | | | | | | |
| Caucasian/White | 79.8 | 74 | | 76.3 | 67.3 | |
| African/Black | 6.1 | 8.3 | | 5.3 | 10.8 | |
| Hispanic/Latino | 3.7 | 6.6 | | 5.8 | 5.4 | |
| East Asian | 0.6 | 2.8 | | 4.3 | 4.9 | |
| South Asian | 1.2 | 2.2 | | 0.0 | 4.9 | |
| Native/Aboriginal | 0.0 | 1.1 | | 1.0 | 0.9 | |
| Middle Eastern | 0.6 | 0.6 | | 0.0 | 0.4 | |
| Other (e.g. mixed-race) | 0.8 | 4.4 | | 7.2 | 4.9 | |
| Education attainment (%) | | | | | | |
| Less than high school | 0.6 | 0.0 | | 0.5 | 0.5 | |
| High school | 4.9 | 8.3 | | 7.7 | 7.2 | |
| Some college | 39.3 | 30.4 | | 39.1 | 36.0 | |
| Bachelor's degree | 33.7 | 36.5 | | 38.2 | 41.4 | |
| Some graduate work | 9.2 | 8.8 | | 1.4 | 2.3 | |
| Advanced degree | 12.3 | 16.0 | | 13.0 | 12.6 | |
| Median income (USD) | 51,000 − 60,999 | 51,000 − 60,999 | | 51,000 − 60,999 | 51,000 − 60,999 | |

*Note*: Standard deviations of means are presented in parentheses. *d* = Cohen's *d*.

modeling (SEM) in Amos 19.0 (Arbuckle, 2010). First, we tested configural invariance between the two experimental conditions (see Vandenberg & Lance, 2000). We tested configural invariance allowing all items of each dimension to load onto the same latent factor for both the experimental and control conditions. For each dimension, the first item s factor loading was fixed to 1.0 while allowing the other items to freely load onto the latent factor and the intercept of the first item was set to 0.0. If this model achieves a good fit, it suggests that the factor structure of the scale fits the data well and respondents across conditions employed similar conceptual factor structure allowing for further tests of measurement invariance. Results for the configural invariance model demonstrated overall acceptable model fit, $\chi^2$ (484) = 939.98, $p < .001$,

TABLE 2
Measurement Invariance Tests of the OCB Scale

| *Models* | *df* | $\chi^2$ | $p(\chi^2)$ | *RMSEA* | *CFI* | $\Delta df$ | $\Delta\chi^2$ | $p(\Delta\chi^2)$ |
|---|---|---|---|---|---|---|---|---|
| Study 1 | | | | | | | | |
| Configural invariance | 484 | 939.98 | <.001 | 0.05 | 0.86 | – | – | – |
| Metric invariance | 503 | 961.38 | <.001 | 0.05 | 0.86 | 19 | 21.40 | 0.32 |
| Scalar invariance | 522 | 993.87 | <.001 | 0.05 | 0.86 | 19 | 32.49 | 0.03 |
| Partial scalar invariance[a] | 521 | 986.34 | <.001 | 0.05 | 0.85 | 18 | 24.96 | 0.13 |
| Equivalent factor means | 526 | 993.59 | <.001 | 0.05 | 0.85 | 5 | 7.25 | 0.20 |
| Study 2 | | | | | | | | |
| Configural invariance | 484 | 941.43 | <.001 | 0.05 | 0.92 | – | – | – |
| Metric invariance | 503 | 961.52 | <.001 | 0.05 | 0.92 | 19 | 20.09 | 0.39 |
| Scalar invariance | 522 | 982.85 | <.001 | 0.05 | 0.92 | 19 | 21.33 | 0.32 |
| Equivalent factor means | 527 | 991.80 | <.001 | 0.05 | 0.92 | 5 | 8.95 | 0.11 |

*Note*: [a] The partial scalar invariance model in Study 1 is tested against the metric invariance model.

RMSEA = .05, CFI = .86 (see Table 2), suggesting that the factor structure and loadings of OCB are satisfactorily equivalent across the two conditions.

Next, we tested metric invariance in which the factor loadings of the same survey items were constrained to be equal between experimental and control groups. This is a strong test of the invariance in factors, which tells us whether or not the same survey item relates to the underlying latent factor the same way between the two conditions. Results for this metric invariance model indicated overall acceptable model fit, $\chi^2$ (503) = 961.38, $p < .001$, RMSEA = .05, CFI = .86 (see Table 2). Critically, the $\chi^2$ difference test between this metric model and the prior (configural) model was non-significant, $\Delta\chi^2 = 21.40$, $\Delta df = 19$, $p > .05$, suggesting metric invariance of the scale across experiment conditions. Results indicated that the OCB scale is structurally similar for participants who are exposed to instructed-response items and those who are not.

To provide an even more stringent measurement invariance test, we conducted a test of scalar invariance. In this test, intercepts of the same survey items were constrained to be equal between experimental and control groups. Although this test is the least frequently conducted test of measurement invariance, some researchers found its results useful for interpreting response threshold differences between groups on the rating of a particular item (see Vandenberg & Lance, 2000). We tested the scalar invariance model of the OCB scale. Results indicated overall acceptable model fit, $\chi^2$ (522) = 993.87, $p < .001$, RMSEA = .05, CFI = .85 (see Table 2). However, the $\chi^2$ difference test between this scalar model and the prior (metric) model was significant, $\Delta\chi^2 = 32.49$, $\Delta df = 19$, $p = .03$, suggesting that not all item intercepts were the same across the two experimental conditions. To identify the source of scalar inequivalence, we examined the item intercepts between the experimental and

control groups (Vandenberg & Lance, 2000). The results indicated that item 16 (see Appendix B), and only this item, had significantly different intercepts across the conditions (experimental = 0.08 vs. control = −0.38). To test for partial scalar invariance, we constrained the intercepts for all survey items to be equal across the conditions except for item 16. Results for the partial scalar invariance model indicated overall acceptable fit, $\chi^2$ (521) = 986.34, $p < .001$, RMSEA = .05, CFI = .85 (see Table 2). Moreover, the $\chi^2$ difference test between the metric and partial scalar invariance models was non-significant, $\Delta\chi^2 = 24.96$, $\Delta df = 18$, $p > .05$. Overall, scalar invariance test results suggested that the item score intercepts of the OCB scale were very similar for respondents who received the instructed-response items and those who did not.[5]

In sum, Study 1 suggests no evidence that instructed-response items are a threat to scale validity. Respondents seeing attention checks or not in the survey did not differ in their responses to and understanding of the scale. Going beyond the current findings, we also wanted to find out if an IMC item poses a scale threat. Compared to instructed-response items, an IMC is more elaborate and can seem trickier to participants, which may induce deliberation and influence a respondent s subjective survey experience more strongly. To test the effect of an IMC on scale responses and replicate the findings, we conducted Study 2.

## STUDY 2

This study tests whether having answered an IMC influences responses to a subsequent scale. Consistent with Study 1, we used the organisational citizenship behavior scale (Podsakoff et al., 1990) as the criterion.

### Method

*Participants and Procedures.* We used the same recruitment and selection procedures as Study 1. Participants completed a 5-minute online survey about workplace experiences for US$0.30. No respondents in Study 2 had participated in Study 1. Survey respondents first reported their demographics and

[5] We also tested equal factor means between groups using structural equation modeling. In this test, the means of the same factor were constrained to be equal between experimental and control groups. This test tells us whether or not there are significant differences between groups on how they scored on each factor of the scale. Results for equal factor means indicated overall acceptable model fit, $\chi^2$ (526) = 993.59, $p < .001$, RMSEA = .05, CFI = .85 (see Table 2). Moreover, the $\chi^2$ difference test between the partial scalar invariance and equal factor means models was non-significant, $\Delta\chi^2 = 7.25$, $\Delta df = 5$, $p > .05$. Consistent with the between-subject ANOVA results, including instructed-response items did not affect respondents' mean scores of the scale.

were then randomly assigned to either the experimental or the control condition. In the experimental condition, respondents answered an IMC (Oppenheimer et al., 2009) (see exact materials in Appendix A) prior to completing the OCB scale. In the control condition, respondents did not answer an IMC. To ensure sufficient power for measurement invariance tests, we aimed to recruit at least 300 participants in total, and collected a final sample of 365. Detailed demographics are reported in Table 1.[6]

## Results

*Mean Differences.* To examine whether the IMC influenced scale responses, we first compared the scale scores across the two conditions. Results of a between-subjects ANOVA indicated that the experimental and control conditions did not differ in overall mean scores or mean scores within each sub-dimension, $ps > .37$ (see Table 1). These results suggest that attention checks did not influence respondents degree of endorsement of items.

*Measurement Invariance.* To assess whether the IMC influenced the way respondents construe the meaning of the items of the scale, we conducted measurement invariance tests using the same procedures as Study 1 (Vandenberg & Lance, 2000). First, we established a baseline model of configural invariance. Results for the configural invariance model indicated overall acceptable model fit, $\chi^2$ (484) = 941.43, $p < .001$, RMSEA = .05, CFI = .92 (see Table 2), suggesting that the factor structure and loadings of the OCB scale are satisfactorily equivalent across the two conditions, which allows for further measurement invariance tests.

Next, we tested metric invariance. Results for the metric invariance model indicated overall acceptable model fit, $\chi^2$ (503) = 961.52, $p < .001$, RMSEA = .05, CFI = .92 (see Table 2). Moreover, the $\chi^2$ difference test between the two models was non-significant, $\Delta\chi^2 = 19.15$, $\Delta df = 19$, $p > .05$, suggesting that the OCB scale is structurally similar for respondents receiving an IMC or not.

[6] In the experimental condition, 21 participants provided an incorrect answer to the IMC. These participants did not differ in overall and sub-dimensional OCB scale scores compared to the participants who successfully completed the attention check and in the control condition. Consistent with Study 1, including these participants or not did not change the overall patterns of results. For a more stringent test of the hypotheses, they were excluded from the analysis. Moreover, there was a lower number of male participants in Study 2 (see Table 1). However, the pattern of our results was consistent across gender, and therefore gender was not included in the main analyses.

To yield an even more stringent test of measurement invariance, we conducted a test of scalar invariance. The intercepts of the same survey items were constrained to be equal between experimental and control groups. Results for the scalar invariance model indicated satisfactory model fit, $\chi^2$ (522) = 982.85, $p < .001$, RMSEA = .05, CFI = .92 (see Table 2). Moreover, the $\chi^2$ difference test between the metric invariance and scalar invariance models was non-significant, $\Delta\chi^2 = 21.33$, $\Delta df = 19$, $p > .05$. Results suggested that the item score intercepts of the OCB scale were similar for respondents who received the IMC and those who did not, supporting that attention checks did not alter scale validity.[7]

## OVERALL DISCUSSION

Taken together, findings from two separate studies both suggest no evidence that attention checks pose a threat to scale validity. Contrary to what the extant literature may have predicted, attention checks did not influence respondents answers to and understanding of the scale. This was consistent regardless of whether the attention checks were in the form of embedded items (Study 1) or as an individual IMC (Study 2). These results contribute to organisational science and other literatures more broadly. To our knowledge, these studies are the first to demonstrate that attention checks do not seem to bear an underlying threat to scale validity. Because attention checks are so widely used nowadays, this is an especially timely piece of evidence. The findings also contribute to our understanding of survey methods and help justify researchers use of attention checks to ensure quality data. Moreover, as the wording of instructed-response items and IMCs are very similar across studies in the literature, our findings can generalise to many other attention check variations. One variation would be the increasingly popular "infrequency items"—questions that yield an obvious logically right answer.[8] Resembling IMCs and instructed-response items, infrequency items may increase deliberation, but our research would suggest that they should not pose scale validity concerns. Furthermore, even though the studies have a strong focus on a management science audience, our studies are just as important in informing scholars in other academic fields that frequently use survey designs, such as psychology, education, political science, and communication studies.

---

[7] We also tested equal factor means between groups using structural equation modeling. Results indicated satisfactory model fit, $\chi^2$ (527) = 991.80, $p < .001$, RMSEA = .05, CFI = .92 (see Table 2). Moreover, the $\chi^2$ difference test between the partial scalar invariance and equal factor means models was non-significant, $\Delta\chi^2 = 8.95$, $\Delta df = 5$, $p > .05$. Consistent with the between-subjects ANOVA results, including instructed-response items did not affect respondents' mean scores of the scale.

[8] "I work fourteen months in a year" (Huang, Bowling, Liu, & Li, 2015).

Theoretically, our findings also add to the literature on deliberation and bias, and generate interesting future directions for unpacking the null effect. Initially, we grounded our hypothesis on the prior literature; because attention checks increase deliberation (Hauser & Schwarz, 2015), and deliberation can cause inaccurate or inconsistent responses (e.g. Dijksterhuis et al., 2006; Nordgren & Dijksterhuis, 2009), attention checks would bias survey responses and threaten scale validity. But our data found no support for this claim and could not reject the null hypothesis—attention checks did *not* influence people's scale responses. Notably, one should be very hesitant in interpreting null findings. Yet, if we entertain the possibility that the null hypothesis is true, future research would benefit from generating new ideas to reconcile the discrepancy between our findings and the theoretical argument. For example, why do attention checks increase deliberation enough to affect cognitive task performance (Hauser & Schwarz, 2015), but not enough to affect survey responses? Are survey responses simply not vulnerable to deliberation effects (cf. Schwarz, 1994)? Or perhaps, even though attention checks generally do not affect scale validity, there could be boundary conditions where attention checks do alter scale responses (e.g. depending on scale characteristics, individual differences of survey respondents). These are interesting possibilities and empirical questions for future research to explore.

Thus far, even though the article has emphasised the potential downside of including attention checks, one should not undermine the benefits of employing attention checks to ensure quality survey responses. In fact, there are reasons why in some cases using attention checks could be especially beneficial. For instance, when the incentive is low for survey takers to provide careful responses (e.g. no pay, time pressure), careless responding tends to be higher. In this situation, the use of attention checks allows researchers to screen out careless responses effectively, preventing careless responding from damaging scale validity (Meade & Craig, 2012). When the incentive for completing the survey carefully is high (e.g. in job interview and assessment), attention checks may be less necessary. Moreover, attention checks may serve as a warning to careless participants in the survey. Research has shown that giving warnings can effectively reduce careless responses. Explicit warning instructions such as "…responding without much effort will be flagged for low-quality data" (Ward & Pond, 2015) and "…responding without much effort would result in loss of credits" (Huang et al., 2012) were shown to increase quality in responses. Because survey respondents who read the attention check can tell it is a "trap", it signifies that the researchers are flagging low-quality data, a function that resembles what a warning basically does. As a warning, attention checks may deter respondents from carelessly responding in the rest of the survey. This would provide a case arguing *for* the use of attention checks instead. It is because,

through motivating survey takers to respond more carefully, there is potential that attention checks in fact increase—rather than decrease—measurement validity.

Despite broad implications, our studies have limitations. The studies tested the effects of attention checks on one specific scale, and they cannot provide direct evidence to address whether or not these attention checks influence responses to other scales. Although the current tests were meant to be conservative (i.e. we used a multidimensional scale, and the scale occurs with or immediately after the attention check), replication research on scales that have other characteristics would be valuable. For example, the simplicity of a decision can reduce the effect of deliberation on poorer judgment (Dijksterhuis et al., 2006). Building on our prior discussion about possibilities for the null effect, one possible reason why attention checks did not significantly affect the OCB scale responses was that respondents found the items simple and clear. Despite the multiple factors of the scale, many OCB items are about concrete actions (e.g. "does not take extra breaks"), leaving relatively little room for ambiguous interpretation. On the contrary, deliberation would lead to more biased responses when the decision is complex. Attention checks may have a stronger effect altering item responses when scale items are complex, ambiguous, or contain irrelevant information. This also speaks to why a good scale item should be clear and concise; no double-barrels and multiple interpretations.

Moreover, although our study samples are very diverse, one limitation is that they are both recruited from the same source, Amazon s Mturk. Mturk samples are more experienced in completing surveys (Hauser & Schwarz, 2016; cf. Buhrmester et al., 2011; Goodman, Cryder, & Cheema, 2013), and some have argued that experience may influence survey responses in general (Berinsky, Huber, & Lenz, 2012; Chandler, Mueller, & Paolacci, 2014; Transue, Lee, & Aldrich, 2009). From our current studies, we do not know if these Mturkers reaction to attention checks can generalise to other less experienced samples. However, it is noteworthy that studies demonstrated that the effect of attention checks on increased deliberation does not depend on the familiarity of attention checks (Hauser & Schwarz, 2015), which seems to suggest that experience in survey taking may not be an issue. To empirically test the generalisability of the results, nonetheless, future research should replicate the current findings with other samples.

Although attention checks do not seem to influence scale responses in general, some personality differences may contribute to people's varying susceptibility to the effect of attention checks. For instance, past research has shown that some people are in general more suspicious than others (e.g. Couch, Adams, & Jones, 1996). Suspicious people may be influenced by attention checks more strongly because they are more likely to infer that the goal of a

study is "more than meets the eye". In contrast, it is also possible that they could be affected by attention checks less strongly because they are more used to a less trusting environment. The effects of individual differences on survey response style appear to be an interesting avenue to explore.

Furthermore, whereas scale validity could be one direct outcome influenced by attention checks, there could be other and more indirect ways in which attention checks affect the validity of measurements. Take convergent and discriminant validity as an example. By inducing more deliberate thinking, attention checks may alter the way people construe relations between the constructs measured in a survey. If this is true, attention checks will dampen convergent and discriminant validity—the degree to which the focal concept is observed to be similar to related constructs (i.e. convergent) and distinct from unrelated constructs (i.e. discriminant) as theories would have predicted. One way to test this phenomenon is to examine whether attention checks affect how well scale measures fit into their nomological networks (Cronbach & Meehl, 1955). However, because the literature on the effects of attention checks on scale responses is limited, whether attention checks affect other forms of measurement validity still awaits more empirical work. Through further understanding the interplay between individuals and survey characteristics, future research will benefit and continue to improve the quality of survey methods and findings.

## Conclusion

Attention checks have become a popular method in survey design to ensure quality samples and hence the validity of scale measurements. However, very recent evidence suggests that attention checks may influence respondents level of deliberation, causing a potential threat to scale validity, which attention checks are to protect. Our findings provided a critical and timely test and found no evidence that attention checks significantly affected scale responses. Researchers may continue utilising attention checks in survey designs and examining other dynamics between respondents and survey characteristics to advance our research methods in general.

## REFERENCES

Arbuckle, J. (2010). *IBM SPSS Amos 19 user's guide*. Crawfordville, FL: Amos Development Corporation.

Berinsky, A.J., Huber, G.A., & Lenz, G.S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Berinsky, A.J., Margolis, M.F., & Sances, M.W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys.

*American Journal of Political Science*, *58*(3), 739–753. http://doi.org/10.1111/ajps.12081

Berry, D.T.R., Wetter, M.W., Baer, R.A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340–345. http://doi.org/10.1037/1040-3590.4.3.340

Bowling, N.A., Huang, J.L., Bragg, C.B., Khazon, S., Liu, M., & Blackmore, C.E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218–229. http://doi.org/10.1037/pspp0000085

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-qulity, data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://doi.org/10.1177/1745691610393980

Calvillo, D.P., & Penaloza, A. (2009). Are complex decisions better left to the unconscious? Further failed replications of the deliberation-without-attention effect. *Judgment and Decision Making*, *4*(6), 509–517.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïvete among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. http://doi.org/10.3758/s13428-013-0365-7

Couch, L.L., Adams, J.M., & Jones, W.H. (1996). The assessment of trust orientation. *Journal of Personality Assessment*, *67*(2), 305–323. http://doi.org/10.1207/s15327752jpa6702_7

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. http://doi.org/10.1037/h0040957

Curran, P.G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. http://doi.org/10.1016/j.jesp.2015.07.006

Curran, P.G., & Hauser, D.J. (2015). Understanding responses to check items: A verbal protocol analysis. Paper presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Custers, E.J.F.M. (2014). Unconscious thought and deliberation without attention: A miracle or a mirage? *Perspectives on Medical Education*, *3*(3), 155–158. http://doi.org/10.1007/s40037-014-0127-y

DeSimone, J.A., Harms, P.D., & DeSimone, A.J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. http://doi.org/10.1002/job.1962

Dijksterhuis, A., Bos, M.W., Nordgren, L.F., & van Baaren, R.B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, *311*(2006), 1005–1007. http://doi.org/10.1126/science.1121629

Ferguson, E., Matthews, G., & Cox, T. (1999). The Appraisal of Life Events (ALE) scale: Reliability and validity. *British Journal of Health Psychology*, *4*(2), 97–116. http://doi.org/10.1348/135910799168506

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Goodman, J.K., Cryder, C.E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. http://doi.org/10.1002/bdm.1753

Hauser, D.J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*(2), 1–6. http://doi.org/10.1177/2158244015584617

Hauser, D.J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. http://doi.org/10.3758/s13428-015-0578-z

Huang, J.L., Bowling, N.A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299–311. http://doi.org/10.1007/s10869-014-9357-6

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. http://doi.org/10.1007/s10869-011-9231-8

Huang, J.L., Liu, M., & Bowling, N.A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. http://doi.org/10.1037/a0038510

Johnson, J.A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. http://doi.org/10.1016/j.jrp.2004.09.009

Levine, G.M., Halberstadt, J.B., & Goldstone, R.L. (1996). Reasoning and the weighting of attributes in attitude judgments. *Journal of Personality and Social Psychology*, *70*(2), 230–240. http://doi.org/10.1037/0022-3514.70.2.230

Mamede, S., Schmidt, H.G., Rikers, R.M.J.P., Custers, E.J.F.M., Splinter, T.A.W., & van Saase, J.L.C.M. (2010). Conscious thought beats deliberation without attention in diagnostic decision-making: At least when you are an expert. *Psychological Research*, *74*(6), 586–592. http://doi.org/10.1007/s00426-010-0281-8

Maniaci, M.R., & Rogge, R.D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. http://doi.org/10.1016/j.jrp.2013.09.008

Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. http://doi.org/10.1037/a0028085

Nordgren, L.F., & Dijksterhuis, A. (2009). The devil is in the deliberation: Thinking too much reduces preference consistency. *Journal of Consumer Research*, *36*(1), 39–46. http://doi.org/10.1086/596306

Oppenheimer, D.M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. http://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. http://doi.org/10.1177/0963721414531598

Podsakoff, P.M., MacKenzie, S.B., Moorman, R.H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *Leadership Quarterly*, *1*(2), 107–142. http://doi.org/10.1016/1048-9843(90)90009-7

Schmitt, N., & Stults, D.M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373. http://doi.org/10.1177/014662168500900405

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 26, pp. 123–162). San Diego, CA: Academic Press.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105.

Stanovich, K.E., & West, R.F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. http://doi.org/10.1017/S0140525X00003435

Stanovich, K.E., & West, R.F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695. http://doi.org/10.1037/0022-3514.94.4.672

Tordesillas, R.S., & Chaiken, S. (1999). Thinking too much or too little? The effects of introspection on the decision-making process. *Personality and Social Psychology Bulletin*, *25*(5), 625–631. http://doi.org/10.1177/0146167299025005007

Transue, J.E., Lee, D.J., & Aldrich, J.H. (2009). Treatment spillover effects across survey experiments. *Political Analysis*, *17*(2), 143–161. http://doi.org/10.1093/pan/mpn012

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. http://doi.org/10.1177/109442810031002

Van Lange, P.A.M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349. http://doi.org/10.1037/0022-3514.77.2.337

Ward, M.K., & Pond, S.B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554–568. http://doi.org/10.1016/j.chb.2015.01.070

Wilson, T.D., Hodges, S.D., & LaFleur, S.J. (1995). Effects of introspecting about reasons: Inferring attitudes from accessible thoughts. *Journal of Personality and Social Psychology*, *69*(1), 16–28. http://doi.org/10.1037/0022-3514.69.1.16

Wilson, T.D., & Schooler, J.W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*(2), 181–192. http://doi.org/10.1037/0022-3514.60.2.181

Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. http://doi.org/10.1007/s10862-005-9004-7

## APPENDIX A

Instructional Manipulation Check (adapted from Oppenheimer et al., 2009)

**Workplace Facilities**

Most modern theories of psychology recognize the fact that social perceptions do not take place in a social vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the perception process. In order to facilitate our research on perceptions of workplace behaviors, we are interested in knowing certain factors about you, the perceiver. Specifically, we are interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions **will be ineffective. So, in order** to demonstrate that you have read the instructions, please ignore the **facility items below. Instead,** simply click the other option and in the corresponding box, enter the t**ext: I read the instructions.**

**Which of these facilities are available at your workplace?**
(Click on all that apply)

- ☐ Canteen/vending machine
- ☐ Lounge
- ☐ Coffee maker
- ☐ Air conditioning/heating
- ☐ Storeroom
- ☐ Washroom
- ☐ Windows
- ☐ Parking
- ☐ Childcare Facilities
- ☐ Other

## APPENDIX B

Organisational Citizenship Behavior Scale (Podsakoff et al., 1990)

Please rate the extent to which each of the following statements describes you in your workplace.

1. Helps others who have heavy work loads
2. Is the classic "squeaky wheel" that always needs greasing
3. For this question, please select number two to demonstrate your attention*
4. Believes in giving an honest day's work for an honest day's pay
5. Consumes a lot of time complaining about trivial matters
6. Tries to avoid creating problems for co-workers
7. Keeps abreast of changes in the organisation
8. Tends to make "mountains out of molehills"
9. Considers the impact of his/her actions on co-workers
10. Attends meetings that are not mandatory, but are considered important
11. Is always ready to lend a helping hand to those around him/her
12. Attends functions that are not required, but help the company image
13. Reads and keeps up with organisation announcements, memos, and so on
14. Helps others who have been absent

15. Does not abuse the rights of others
16. Willingly helps others who have work-related problems
17. Always focuses on what's wrong, rather than the positive side
18. For this question, please select number six to demonstrate your attention*
19. Takes steps to try to prevent problems with other workers
20. Attendance at work is above the norm
21. Always finds fault with what the organisation is doing
22. Is mindful of how his/her behavior affects other people's jobs
23. Does not take extra breaks
24. Obeys company rules and regulations even when no one is watching
25. Helps orient new people even though it is not required
26. Is one of the most conscientious employees

*Additional instructed-response items that appeared only in the experimental condition, Study 1.