# The Incremental Value of Controlling for Covert Insufficient Effort Responding

**Ann-Marie R. Castille**
**Nicholls State University**

**Christopher M. Castille**
**Nicholls State University**

**Sandesh Sharma**
**Nicholls State University**

*Insufficient effort responding (IER) is a common concern of survey researchers especially those who collect data through crowdsourcing. Methods of controlling for IER may be overt (identifiable by respondents) or covert. This study examines the relative impact of controlling for covert IER when overt-IER methods are in the survey design. Using data from an experiment on performance feedback reactions where overt IER controls were in place, we examine the scale reliabilities and convergent and discriminant validity, both of which change negligibly by controlling for covert IER. Findings suggest controlling for covert IER lacks incremental value beyond controlling for overt IER.*

*Keywords: insufficient effort responding, surveys, screening, crowdsourcing*

## INTRODUCTION

Insufficient effort responding (IER) is a common concern of survey researchers. IER has been shown to harm reliability of measurements, posing a threat to construct and statistical conclusion validity in the social sciences (Breitsohl & Steidelmüller, 2018; Meade & Craig, 2012). This is problematic as data quality problems can cause spurious or misleading associations between measures depending upon its nature and prevalence (Chandler et al., 2020; Huang et al., 2012). Controlling for IER is also particularly relevant in a time where survey researchers in disciplines related to business or psychology rely increasingly on methods – such as crowdsourcing – for hypothesis testing purposes, which may be rife with data quality issues (see Keith et al., 2017; McGonagle, 2015).

Controlling for IER comes in at least two broad varieties: *overt* and *covert* methods (see Good et al., 2019). Overt methods are identifiable by respondents and provide opportunities for demonstrating that one is paying attention while completing the survey (e.g., agreeing to a statement "I am paid bi-weekly by leprechauns" when one is not). Such bogus items help survey designers identify and remove careless respondents from their datasets, thereby guarding against overt IER (Meade & Craig, 2012). Another

method of overt IER detection is the *factual manipulation check*, which is an objective question about a study's key components (Kane & Barabas, 2019).

Covert methods are unidentifiable by respondents and involve post hoc analyses conducted after data collection is complete (Meade & Craig, 2012). Such methods involve using global survey response behavior (to identify IER. Overt and covert methods appear to reflect largely unrelated forms of survey misbehavior, leading scholars to suggest that both overt and covert methods are necessary for data cleaning (DeSimone & Harms, 2018; Meade & Craig, 2012).

What has not been examined is the relative impact of controlling for covert forms of IER when overt IER methods are utilized in the survey design process. Overt forms of IER can with relative ease be controlled in the design of a survey (e.g., if a response is flagged with an overt IER check, participants can be warned or removed from the survey mid-administration). By contrast, covert forms of IER will often require more time in the data cleaning phase to identify any IER that may be missed by overt methods (Meade & Craig, 2012). What remains unclear is whether ignoring covert forms of IER is as unwise as scholars generally assume (see DeSimone & Harms, 2018).

In our study, we examine the incremental value of controlling for covert forms of IER when overt methods are involved in the survey design. We present data where overt forms of IER were controlled for in the design phase of an experiment examining substantive phenomena of interest to business and psychology professionals (e.g., goal setting, performance, affective reactions to performance feedback). We highlight how measurement reliability (i.e., indicated via Cronbach's alpha) changes as a function of controlling for covert IER. Further, we highlight how effect sizes change as additional covert IER screening measures are implemented. Whereas other scholars have examined the relationship between overt and covert IER strategies (e.g., DeSimone & Harms, 2018; Good et al., 2019), we focus on the impact of screening for covert IER when overt methods are utilized.

## STUDY CONTEXT AND RESEARCH QUESTION

The data for our investigation come from a large experiment involving working individuals who were sampled using Amazon's Mechanical Turk (MTurk). In this experiment, we examined affective and cognitive reactions to performance feedback as well as goal-setting behavior. Overt forms of IER were identified and removed either during the survey administration or post-hoc via a manipulation check that was relevant for the study. The manipulation check involved asking participants to report whether they had received positive or negative feedback that was either nominal (e.g., you did/did not meet a stated personal goal) or relative (e.g., you did/did not exceed the performance of rivals). Individuals who could not recall or reported feedback interventions that were not assigned to them were ejected from the study. This can be seen as a form of controlling for IER. To examine the incremental value of controlling for covert forms of IER, we employed a measure of personal reliability/response consistency and survey duration. Both methods flag individuals for covert forms of IER and are widely used in the literature (e.g., DeSimone & Harms, 2018; Good et al., 2019). Additionally, after controlling for covert forms of IER, we re-examined estimates for internal consistency and scale correlations to determine if any substantive changes in conclusions would necessarily follow. Such activities help to answer one key question raised by our study:

***Research Question:*** *Does controlling for covert forms of IER after overt forms of IER have been controlled produce different estimates of internal reliability and scale validity?*

Scholars appear unanimous that data quality can be a concern for MTurk. As this experiment drew upon MTurk's population, it provides a useful context for examining the impact of controlling or not controlling for covert IER. Furthermore, as the substantive phenomena that we have examined (e.g., reactions to performance feedback) are well-studied in the literature, we also highlight relevant estimates from meta-analyses and a primary study on scale reliability and validity where they are available. What this comparison enables us to discuss are study-specific causes of measurement error that can be attributed to IER. Given that the prevalence of IER within a study varies due to a variety of factors (see Huang et al., 2012; DeSimone

& Harms, 2018) – i.e., IER should not be linked to any specific method, measure, or construct – it may be considered a source of measurement error (e.g., transient, situation-specific) from a generalizability theory perspective. In other words, uncorrected meta-analytic estimates, particularly for scale reliability and validity (e.g., correlations with other scales), provide a useful comparison to our own estimates as these would be unlikely to be affected by IER.

## METHOD

### Sample

The sample included participants from MTurk who were native English speakers and at least 18 years of age living in the U.S. or Canada. Data was gathered from a total of 1,027 respondents as part of a larger study with 848 respondents providing complete data. The sample demographics were fairly representative of an MTurk study conducted by I/O psychologists (for comparisons, see Keith et al., 2017): mean age was 33.84 (SD=10.17); the sample was female-dominated (i.e., 54.1%); the majority of respondents were Caucasian (80.31%) followed by Asian (7.78%), and African American (6.96%).

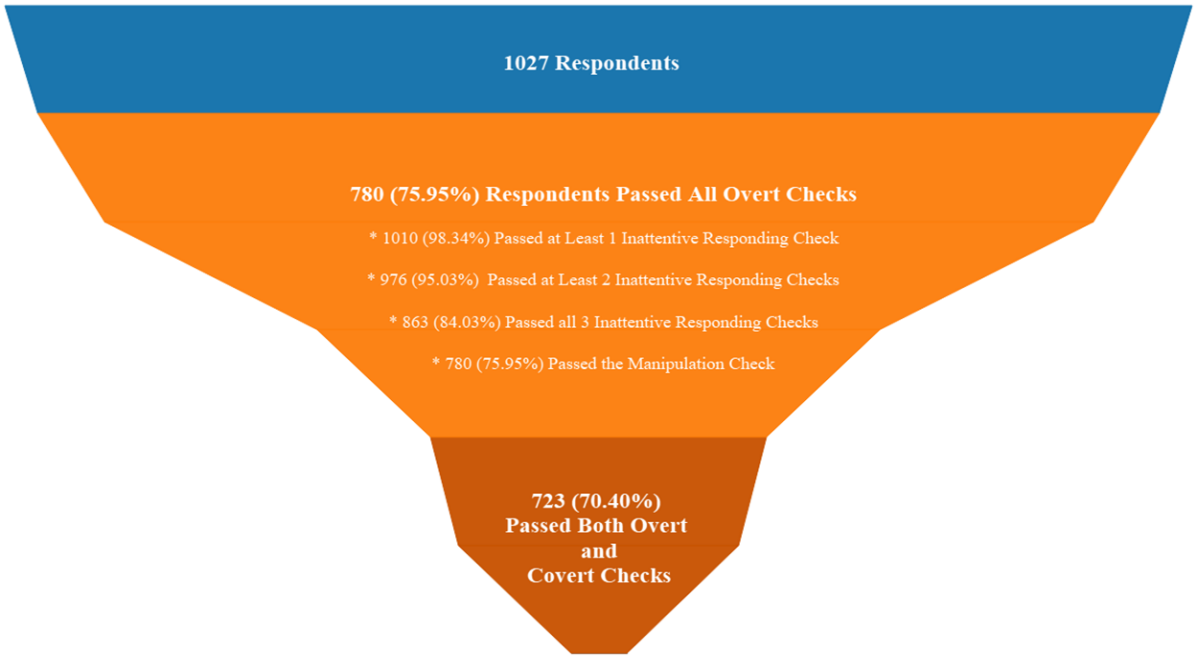
### Procedure

Participants were trained to perform a complex task: conduct an online search to identify RGB codes for a specified university's primary colors. Afterward, participants were asked to set a performance goal and given three minutes to provide RGB codes for a maximum of six specified universities. Following the task, they were shown a feedback message and proceeded with a survey on their reactions to the feedback. They then set a goal and performed a second round task, and finally completed a manipulation check and demographic questionnaire.

### Measures

To assess IER, we used both overt and covert indicators, which are explained below. See Figure 1 for a visual depiction of filtering for overt and covert IER.

**FIGURE 1**
**PASSING RATES FROM EACH IER SCREEN**



*Overt IER*

Overall, 247 respondents were screened out from overt checks which included three randomly placed bogus survey items (Meade & Craig, 2012) and two manipulation check items. The bogus survey items resulted in 164 respondents being screened out. The manipulation check resulted in an additional 83 respondents being screened out.

*Covert IER*

A total of 57 respondents were screened out from covert checks which included survey response time and personal reliability. The covert analysis was only conducted using the data of respondents who passed all overt checks. For response time, we examined whether any respondents were outliers and considered negative outliers as failing the check; however, no negative outliers existed. Personal reliability was assessed using split-half correlations (corrected using Spearman-Brown Prophecy Formula) among the halves of the survey items. Respondents with a personal reliability below 0.3 failed the covert IER check (Johnson, 2005).

*Reactions to Performance Feedback*

The measures included in the survey were the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988) and Leventhal's (1976) distributive justice (DJ) measure, both of which have demonstrated high reliability and validity.

*Goal Setting and Performance*

Goal setting was operationalized as the goal level set by participants for Task 2. Participants were either asked to set a goal for the number of universities for which they would provide RGB codes or for the percentage of participants they would outperform. There were seven goal level options; therefore, their set goal was recorded as one through seven. Performance level was measured as the number of universities for which participants correctly provided RGB codes. The maximum possible performance level was 6.

**Analytical Approach**

First, we identified the participants who failed the overt checks and calculated the proportion of those who passed each check (see Figure 1). After removing these participants, we then calculated (i) Cronbach's alphas for our Likert scales, (ii) the correlation coefficients (Pearson's *r*) between our measures, and (iii) confidence intervals (CIs) for the prior estimates. Then, we removed respondents who failed our covert checks (i.e., response time and personal reliability) and re-estimated our Cronbach's alphas, correlation coefficients, and CIs. Next, we examined whether the CIs of the first overt-screened group overlapped with those of the second covert-screened group. If the CIs overlap, we would conclude that adding in the covert screens adds no incremental value to the data screening process. If the CIs do not overlap, we would conclude that adding in the covert screens does significantly alter the study's findings and may be beneficial. This approach of comparing two samples that include many of the same participants does violate the assumption of independence but is consistent with a strategy of controlling for a contaminating factor and reporting the results of controlling and not controlling for this potential contaminant (see Spector & Brannick, 2011). Lastly, to help our readers interpret our findings within the broader literature, we provide select effect size estimates from the literature (i.e., Cronbach's alpha and correlations, both with CIs).

**RESULTS**

Descriptive statistics for study variables can be found in Table 1. It is worth noting that we only provide statistics for personal reliability as it was redundant with response time in our dataset for screening purposes. As can be seen, personal reliability was positively associated with many of our Likert measures, suggesting that covert IER may be a common contaminant for these measures.

**TABLE 1**
**DESCRIPTIVE STATISTICS FOR STUDY VARIABLES**

| Variable | Mean | SD | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| 1. Personal Reliability | 93.28% | .25 | | | | | |
| 2. Positive Affective State | 2.78 | .97 | .10** | | | | |
| 3. Negative Affective State | 1.57 | .67 | .30*** | .10** | | | |
| 4. Distributive Justice | 3.46 | 1.13 | .23*** | .33*** | .18*** | | |
| 5. Goal Setting | 4.24 | 1.53 | .04 | .16*** | .12*** | .17*** | |
| 6. Task Performance | 2.19 | 1.45 | .02 | .04 | .03 | .07* | .26*** |

Note. These values were calculated using data from 774 individuals who passed all 3 overt IER checks. We only provide descriptive statistics for personal reliability (% passing) because all remaining participants passed the other covert check (i.e., survey duration).

**Scale Reliability as a Function of Controlling for Overt and Covert IER**

Scale reliability estimates for our Likert scales (i.e., Cronbach's alpha) before and after controlling for covert forms of IER are depicted in Table 2. Across all three measures (PA, NA, DJ), controlling for covert forms of IER had a negligible effect on scale reliability (i.e., all CIs overlapped, all changes were trivially small in magnitude). Comparing these estimates to that obtained in the broader literature suggests that our measures performed more reliably regardless of controlling for covert IER.

**TABLE 2**
**SCALE RELIABILITY ESTIMATES BEFORE AND AFTER COVERT SCREEN ALONG WITH PUBLISHED BENCHMARKS**

| Scales | Before Covert Screen | | | After Covert Screen | | | Published Benchmark from the Literature | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 95% CI | | | 95% CI | | | 95% CI | | |
| | $\alpha$ | LL | UL | $\alpha$ | LL | UL | $\alpha$ ($n$) | LL | UL | Source |
| PA | .92 | .92 | .93 | .93 | .92 | .94 | .85 (5651) | .85 | .86 | Leue et al. (2010) |
| NA | .89 | .87 | .90 | .89 | .87 | .90 | .79 (5332) | .78 | .80 | Leue et al. (2010) |
| DJ | .93 | .92 | .94 | .93 | .92 | .94 | .91 (32864) | .91 | .91 | Hauenstein et al. (2001) |

*Note*: PA = Positive Affect; NA = Negative Affect; DJ = Distributive Justice; CI= confidence interval; LL = lower limit; UL = upper limit.

**Select Correlations as a Function of Controlling for Overt and Covert IER**

We report select correlations among our Likert and single-item scales (i.e., goal setting, performance) in table 3. We highlight these correlations as we were able to identify meta-analytic benchmarks for these effects in the literature. We start with the PA-DJ correlation as both of these scales were correlated with our covert IER measure, which indicates a common source of contamination (Spector & Brannick, 2011). Interestingly, the PA-DJ correlation weakens negligibly after controlling for covert IER ($r_{before}$ = .33, 95% CI = .27, .40; $r_{after}$ = .31, 95% CI = .24, .37). Both of our estimates (i.e., before and after controlling for covert IER) produce findings that overlap with a published benchmark ($r_{benchmark}$ = .34, 95% CI = .31, .37). Thus, controlling for covert IER did not alter our conclusions.

Next are the correlations with NA (i.e., NA-PA, NA-DJ), which are measures that are also correlated with our covert IER measure. The NA-PA correlation strengthens negligibly after controlling for covert IER ($r_{before}$ = -.10, 95% CI = -.17, -.03; $r_{after}$ = -.11, 95% CI = -.18, -.04). By contrast, the NA-DJ correlation weakens negligibly ($r_{before}$ = -.180, 95% CI = -.25, -.11; $r_{after}$ = -.177, 95% CI = -.25, -.11). All of our estimates are consistently weaker than published benchmarks (NA-PA $r_{benchmark}$ = -.52, 95% CI = -.55, -.49; NA-DJ $r_{benchmark}$ = -.32, 95% CI = -.30, -.34). We caution against attributing this observation to IER strategy given the fact that our study did not capture much variability in NA (M = 1.57, SD = .97). In other words, range restriction – not IER – appears to have affected our ability to reliably estimate these effects.

The correlation between DJ and performance strengthens negligibly after controlling for covert IER ($r_{before}$ = .07, 95% CI = .00, .14; $r_{after}$ = .08, 95% CI = .01, .16). These estimates are weaker than published benchmarks ($r_{benchmark}$ = .19, 95% CI = .17, .21). Interestingly, controlling for covert IER does produce an estimate that, though negligibly weaker, is nevertheless closer to the published benchmark. It remains unclear why this is the case, suggesting that some unknown study-specific explanation is tenable.

Lastly, the correlation between goal setting and performance weakens negligibly after controlling for covert IER ($r_{before}$ = .26, 95% CI = .19, .32; $r_{after}$ = .24, 95% CI = .17, .31). Interestingly, our estimates are stronger than a published benchmark ($r_{benchmark}$ = .14, 95% CI = .03, .25) but the CIs do overlap. As covert IER was not a contaminant of these measures in our dataset, it remains unclear why this is the case, suggesting that some unknown study-specific explanation is tenable.

Overall, in not one comparison did controlling for covert IER alter our conclusions, nor did controlling for covert IER necessarily allow us to produce estimates that were consistent with the literature. Rather, differences appear due to study-specific factors (e.g., range restriction, other unknown factors). Cumulatively, our findings suggest that controlling for covert IER may not substantially alter study conclusions.

**TABLE 3**
**SELECT CORRELATIONS AND COFIDENCE INTERVAL ESTIMATES BEFORE AND AFTER COVERT SCREEN ALONG WITH PUBLISHED BENCHMARKS**

| Scales | Before Covert Screen | | | | After Covert Screen | | | | Published Benchmark from the Literature | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 95% CI | | | | 95% CI | | | | 95% CI | | |
| | *r* | LL | UL | *p* | *r* | LL | UL | *p* | *r* (*n*) | LL | UL | Source |
| PA & DJ | .33 | .27 | .40 | <.01 | .31 | .24 | .37 | <.01 | .34 (2678) | .31 | .37 | Colquitt et al. (2013) |
| NA & PA | -.10 | -.17 | -.03 | .01 | -.11 | -.18 | -.04 | .01 | -.52 (2298) | -.55 | -.49 | Colquitt et al. (2013) |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NA & DJ | -.18 | -.25 | -.11 | <.01 | -.18 | -.25 | -.11 | <.01 | -.32 (5447) | -.30 | -.34 | Colquitt et al. (2013) |
| DJ & Perf | .07 | .00 | .14 | .04 | .08 | .01 | .16 | .03 | .19 (11336) | .17 | .21 | Colquitt et al. (2013) |
| GS & Perf | .26 | .19 | .32 | <.01 | .24 | .17 | .31 | <.01 | .14 (308) | .03 | .25 | Payne et al. (2007) |

Note. PA = Positive Affect; NA = Negative Affect, DJ = Distributive Justice; GS = Goal Setting; Perf = Performance; FV = Feedback Valence ; CI= confidence interval; LL = lower limit; UL = upper limit. Estimates for (i) PA & NA, (ii) PA & DJ, (iii) NA & DJ, and (iv) DJ & Perf are taken from Colquitt et al.'s (2013) (i) Table 16, (ii and iii) Table 15, and (iv) Table 8 respectively. The estimate for GS & Perf comes from Payne et al.'s (2007) Table 5 (specifically, the correlation between state prove performance goal orientation and task performance).

## DISCUSSION

In examining the incremental impact of controlling for covert IER, specifically personal reliability, we found evidence that covert IER significantly correlated with several of our substantive measures of interest (i.e., PA, NA, DJ). While this did suggest that covert IER contaminated these measures, we did not find any evidence that controlling for covert IER resulted in substantially different (i) scale reliability estimates or (ii) estimates of correlation among our scales. Regarding scale reliability, our estimates for Cronbach's alpha were consistently higher than the meta-analytic literature would otherwise suggest. Since crowdsourced investigations typically produce scale reliabilities that are comparable to conventionally-sourced samples (Walter et al., 2019), this finding may be attributed to our survey design, which omitted individuals who failed overt measures of IER. Regarding estimates of correlation, we found some evidence that correcting for overt IER is sufficient (i.e., covert is unnecessary) to estimate correlations that are well-established in the literature (e.g., the correlation between PA and DJ). Controlling for covert IER did not substantially alter our estimates of correlation. Collectively, these findings suggest that controlling for covert forms of IER may not be necessary once overt methods are utilized.

### Implications

Screening for data quality is often done by diligent researchers to ensure respondents are not engaging in IER. There are many ways to screen for IER (Meade & Craig, 2012), and scholars must make the choice of which method to use. IER researchers have urged scholars to control for both overt and covert forms of IER (see Good et al., 2019; Meade & Craig, 2012; DeSimone & Harms, 2018), however, no one has yet examined the incremental impact of controlling for covert forms of IER once overt forms of IER are controlled by design. This study provides guidance for scholars in making this decision and supports the use of bogus survey items and manipulation checks over survey response time and personal reliability. We contribute to the literature by demonstrating the lack of incremental value of adding the covert IER checks. Our observations also indicate that scholars examining various IER detection and correction practices should examine the relative merits of both overt and covert forms of IER detection in their data cleaning practices.

### Limitations and Future Research Directions

While this research provides valuable contributions to the literature, there are limitations worth noting which will hopefully inspire future research. First, respondents were ejected from the survey if they failed the initial overt IER checks (i.e., three bogus survey items); therefore, we did not gather their data for the manipulation check, survey response time, or personal reliability and could not compare the rates of failure of each approach nor examine the incremental value of adding each check in various orders, both of which would be beneficial areas of future research. A second limitation is that we only used two covert checks,

one of which flagged no additional IER. We encourage researchers to examine alternative covert checks and further assess whether covert checks provide incremental value at identifying IER beyond overt checks. We also encourage future research on the number of screens necessary to ensure that data are sufficiently cleaned for hypothesis testing purposes. Another useful area of research would involve examining the drivers of overt and covert IER, which can inform survey design practices. General audiences would benefit from understanding which practices can prevent IER from occurring.

## CONCLUSION

In the present investigation, we examined whether screening for covert IER, rather than only overt IER, results in a change in the reliability estimates and effect sizes closer to published benchmarks. Since the CIs of our Cronbach's alphas and correlation coefficients between the overt-screened IER responses versus the overt- and covert-screened IER responses overlapped, we conclude that screening for covert IER provides no incremental value beyond simply screening for overt IER. Survey researchers should consider this finding when deciding on their data screening approach for IER.

## REFERENCES


Breitsohl, H., & 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), 284–308. https://doi.org/10.1111/apps.12121

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. *Journal of Abnormal Psychology*, *129*(1), 49–55. https://doi.org/10.1037/abn0000479

Colquitt, J.A., Scott, B.A., Rodell, J.B., Long, D.M., Zapata, C.P., Conlon, D.E., & Wesson, M.J. (2013). Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. *Journal of Applied Psychology*, *98*(2), 199–236. https://doi.org/10.1037/a0031757

DeSimone, J.A., & Harms, P.D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, *33*(5), 559–577. https://doi.org/10.1007/s10869-017-9514-9

Good, S.C., Royes, J.O., & Fisher, D.M. (2019). *Identifying and preventing insufficient effort responding in MTurk samples*. Poster session presented at the 34th annual meeting of the Society for Industrial and Organizational Psychology, National Harbor, MD.

Hauenstein, N.M.A., McGonigle, T., & Flinder, S.W. (2001). A meta-analysis of the relationship between procedural justice and distributive justice: Implications for justice research. *Employee Responsibilities & Rights Journal*, *13*(1), 39–56. https://doi-org.ezproxy.nicholls.edu/10.1023/A:1014482124497

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.

Johnson, J.A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*, 103.

Kane, J.V., & Barabas, J. (2019). No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *American Journal of Political Science*, *63*(1), 234–249.

Keith, M.G., Tay, L., & Harms, P.D. (2017). Systems perspective of amazon mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01359

Leue, A., & Lange, S. (2011). Reliability generalization: An examination of the Positive Affect and Negative Affect Schedule. *Assessment*, *18*(4), 487–501.

Leventhal, G.S. (1976). The distribution of rewards and resources in groups and organizations. In L. Berkowitz & W. Walster (Eds.), *Advances in experimental social psychology* (Vol. 9. pp. 91–131). New York: Academic Press.

McGonagle, A.K. (2015). Participant motivation: A critical consideration. *Industrial and Organizational Psychology*, *8*(2), 208–214.
Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085
Payne, S.C., Youngcourt, S.S., & Beaubien, J.M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology*, *92*(1), 128–150. https://doi.org/10.1037/0021-9010.92.1.128
Spector, P.E., & Brannick, M.T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, *14*(2), 287–305. https://doi.org/10.1177/1094428110369842
Walter, S.L., Seibert, S.E., Goering, D., & O'Boyle, E.H. (2019). A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology*, *34*(4), 425–452. https://doi.org/10.1007/s10869-018-9552-y
Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of a brief measure of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.