What's in a Sample? Comparison of Effect Size Replication and Response Quality across Student, MTurk, and Qualtrics Samples[1]

Kurt Kraiger,[a]* Alyssa K. McGonagle,[b] and Diana R. Sanchez[c]

[a]*Department of Management, University of Memphis, Memphis, United States of America;* [b]*Department of Psychology, University of North Carolina Charlotte, Charlotte, United States of America;* [c] *Department of Psychology, San Francisco State University, United States of America*

*Department of Management, Fogelman College of Business & Economics, 3675 Central Ave., University of Memphis, Memphis, TN 38152 USA. OF: 901.678.3159; email: Kurt.Kraiger@memphis.edu

[1] Paper presented at the 11th Conference on Organizational Psychology: People and Risks, Saratov State University, April 24, 2020

Abstract

Researchers are increasingly using crowdsourcing, or paid study recruitment of participant samples online. A key consideration samples is understanding to what extent results will be generalizable to broader populations. Our study employs three samples: (1) university students, (2) MTurk participants, and (3) a Qualtrics panel. Using these sources, our research makes three contributions to the understanding of how sampling affects the generalizability of research findings. We identify differences across samples in the extent to which they are representative of the U.S. labor force. We examine how well results from each sample replicate a known, meta-analytically determined population effect. We compare samples on four measures of response quality (*Mahalanobis Distance*, *Person-Total Correlations*, *Individual Response Variability*, and *Psychometric Synonyms*) and demonstrate how removing careless responders influences the extent to which sample results replicate population effects. Implications for choosing samples and for the use and reporting of response quality measures are discussed.

Keywords: Crowdsourced samples, survey research, insufficient effort responding, replication, organizational commitment, withdrawal intentions

What's in a Sample? Comparison of Effect Size Replication and Response Quality across Student, MTurk, and Qualtrics Samples

Researchers are increasingly using crowdsourced samples for data collection from providers like Mechanical Turk (MTurk) or Qualtrics (Cheung, Burns, Sinclair, & Sliter, 2017). Crowdsourced participants are typically a demographically diverse, global population (Cheung et al., 2017). However, the validity of data from crowdsourced samples is not yet well-established (e.g., Paolacci & Chandler 2014). Central to the external validity of findings from crowdsourced samples is whether they are representative of the population of interest (Cheung et al., 2017; Landers & Behrend, 2015). Researchers wonder, "Are participants drawn from this sample representative of the population to which I want to generalize?" and "Can I achieve reasonable estimates of the population effect?" Interest in data from crowdsourced samples has intensified of late, evidenced by numerous publications on insufficient effort responding (IER) or careless responding (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Huang, Liu, & Bowling, 2015; McGonagle, Huang, & Walsh, 2016). We investigate these issues by examining whether findings obtained from crowdsourced samples replicate meta-analytic population values and presenting information about how IER affects conclusions drawn from these findings. Specifically, we examine whether results attained from three different sources (college students, MTurk, Qualtrics panel) replicate known population values established by meta-analysis. Additionally, we show that including or removing of participants demonstrating IER affects obtained estimates.

Our study makes two contributions. First, we use a novel indicator of data quality to examine the external validity of crowd-sourced findings –comparison of observed construct correlations with those attained from a meta-analysis. Meta-analyses provide a good reference due to their large sample sizes, stable results, and diverse sample composition. Here, we compare a meta-analytic estimate to correlations between two well-established constructs (organizational commitment and withdrawal intentions) from three samples.

Second, we present and compare results with and without removing participants demonstrating IER. In doing so, we show how comparing meta-analytic findings to sample correlations may be problematic depending on how IER is treated.

Sample Sources: Replication

MTurk is an increasingly popular crowdsourced sample source (Buhrmeister, Kwang, & Gosling, 2011; Paolacci & Chandler, 2014). Survey platforms such as Qualtrics and SurveyMonkey also provide researchers with access to study panels but have been studied less frequently.

The validity of inferences made from crowdsourced samples dates back to Burhmeister et al.'s (2011) seminal "how to" article introducing MTurk to the social sciences. The researchers administered personality questionnaires of different lengths to both MTurk and student samples and compared test reliabilities between samples, concluding "data obtained are at least as reliable as those obtained via traditional methods" (p. 3).

Scores can be reliable but not valid. An early approach to establishing validity was demonstrating data collected from MTurk samples replicated classic effects found in disciplines such as behavioral economics and decision-making (e.g., Paolacci, Chandler, & Ipeirotis, 2010). For example, Paolacci et al. tested several empirically-established decision-making errors in student and MTurk samples, and found effects replicated across samples.

Multiple experimental studies have attempted replication of known effects with crowdsourced samples. Crump, McDonnell, and Gureckis (2013) attempted to replicate a "diverse body of tasks from experimental psychology" using an MTurk sample and reported, "…the data seem mostly in line with a laboratory results so long as the experiment methods were solid" (p. e57410), suggesting that experimental effects found face-to-face can be replicated using crowdsourced samples. Mullinix, Leeper, Druckman, and Freese (2015) found no significant differences in the effects of framing participants' opinions across four samples including MTurk and student samples.

Studies comparing the validity of inferences from traditional samples to crowdsourced samples have nearly all based conclusions of generalizability on statistical significance. This criterion for determining sample estimate equivalence is problematic for two reasons. First, because of their convenience and lower procurement costs, MTurk samples are often considerably larger than traditional samples. An experimental manipulation may be significant in the crowdsourced sample simply because of greater power. Second, behavioral scientists are

consistently encouraged to forego significance testing for effect sizes (e.g., Cumming, 2014). Cumming and others argue that our primary goal for individual studies should be the determination of effect sizes for phenomena of interest.

With respect to survey research and correlational designs, Walter, Seibert, Goering, and O'Boyle (2018) recently conducted a meta-analysis in which they identified 43 correlations between two variables from the organizational sciences (e.g., neuroticism and job satisfaction) in a crowdsourced sample for which there was a meta-analytic estimate. Eighty-six percent of the correlations fell within the 80% credibility intervals of its meta-analysis. While this increases confidence that correlation estimates from crowdsourced samples replicate known effect sizes, Walter et al. did not provide comparisons across types of online samples (e.g., MTurk v. Qualtrics), nor comparisons for student samples.

To date, no studies have compared multiple crowdsourced samples on their effect sizes or compared these effect sizes to those obtained from other samples. Here we examine the extent to which factor correlations between organizational commitment (OC) and withdrawal intentions (WI) from three samples – MTurk, Qualtrics, university students – replicate a meta-analytic estimate of the population effect size. We thus ask:

RQ1: To what extent do data from Student, MTurk, and Qualtrics samples replicate a meta-analytically-established correlation between OC and WI?

IER and Sample Sources

Investigations of IER have increased recently, while introducing new metrics for identifying problematic respondents (e.g., Curran, 2016; Meade & Craig, 2012). Researchers have found IER evidence in crowdsourced samples (e.g., Huang and colleagues, 2012; 2015). Furthermore, there have been numerous recent calls in for researchers to detect and possibly IER (e.g., Huang et al., 2012; Meade & Craig, 2012), and to report transparently their methods for doing so (e.g., McGonagle et al., 2016).

Recognition of IER has implications for interpreting meta-analytic estimates. We assume it is likely that most samples in existing meta-analyses include *unidentified* careless responders as these investigations aggregate primary studies in which sample data were collected before identifying/eliminating cases for IER was common. Therefore, care must be taken when interpreting meta-analysis results as "population coefficients." To demonstrate this, we present sample correlations with and without careless responders and compare both sets of sample correlations to a meta-analytic referent. Because IER inflates observed correlations in certain situations (e.g., when variable means in a bivariate relationship depart from scale midpoints; Huang et al., 2015) and attenuates them in others (due to increased measurement error; McGrath, Mitchell, Kim, & Hough, 2010), we do not consider directionality here. Instead, we propose that such changes may influence conclusions drawn regarding generalizability when using meta-analytic correlations as referents:

RQ2: Will removing careless responders affect conclusions drawn about data quality when using meta-analytic correlations as referents?

Detecting IER. Efforts to detect and control for IER can be taken either before or after data collection (see Curran, 2016; Meade & Craig, 2012). Proactive approaches for detecting IER are discussed in detail in Curran (2016), and Meade and Craig (2012). Post-hoc approaches screen responses after data are collected. Meade and Craig identified multiple classes of post-hoc approaches including outlier analysis, and response consistency. Here we use four post-hoc approaches: *Mahalanobis Distance* (*MD*) and *Person-Total Correlations* (*PTr*) as outlier analyses, plus *Individual Response Variability* (*IRV*) and *Psychometric Synonyms (PS)* for response consistency. Collectively, these four allow us to compare IER detection within and across metrics.

Method

Participants and Procedure. All data were archival. The Student sample consisted of 382 psychology students from a United States public university who completed measures for course credit as part of a broader study on the relationship between job-fit and employee attitudes.

The MTurk sample consisted of 581 U.S. respondents recruited by posting a task for a study examining relationships between job-fit and employee attitudes. MTurk workers were compensated $.75USD for participating (without confirming data quality.

The Qualtrics sample consisted of 321 respondents. Qualtrics distributed the survey link and returned completed data. Qualtrics contracted for a payment of $25USD per participant, of which about $5USD went to each participant.

Measures

Organizational Commitment. OC was measured by the 15-item Organizational Commitment Questionnaire (OCQ; Mowday, Steers, & Porter, 1979), consisting of nine positively worded items and six negatively worded items with a seven-point response format. Alphas were .87 (Student), .93 (MTurk), and .88 (Qualtrics).

Withdrawal Intentions. WI was measured with Mobley, Horner, and Hollingsworth's (1978) three-item measure with a five-point response format. Alphas were .88 (Students), .94 (MTurk), and .95 (Qualtrics).

Analyses

Response Quality. We quantified response quality based on outlier analysis (*MD, PTr*) and response consistency (*IRV*, *PS*). We computed Mahalanobis or *MD* scores for each respondent for both OC and WI scores, and tested for statistical significance to identify potential IER respondents (Meade & Craig, 2012). The distance score represents the multivariate distance between participants' response vectors and the sample mean vector and is useful for detecting individual multivariate outliers (Tabachnick & Fidell, 2007). There are no validated cut-off values for determining IER as outliers. To establish a criterion for IER, we set $p <$ .05 for statistical significance for the chi-square value (Tabachnick & Fidell, 2007),

but used the Holm-Bonferroni correction (Holm, 1979) to control for experiment-wise error rate. Participants with statistically significant distance scores received a "1," indicating IER.

Person-Total Correlations (*PTr*) measure the association between respondents' scores on 15 OC and three WI items with the total or mean score for the sample on the same items (Curran, 2016). Curran recommended negative correlations as a conservative flag for IER. Individuals with negative *PTr* were coded "1" indicating IER; all others were coded "0."

Individual response variability (*IRV*) is a measure of response consistency and detects IER by determining low variability in a set of responses (Dunn, Heggestad, Shannock, & Theilgard, 2018). It is calculated as the standard deviation of responses across consecutive item responses for participants. We calculated *IRV* across nine positive and six negative (prior to reversing) OC items. Low variability across items before reverse scoring negative items reveals IER and higher variability is indicative of less IER. As recommended by Dunn et al., we created a cut-off to identify IER responders by calculating the tenth percentile of *IRV* scores (across samples). The tenth percentile was 0.798; participants with $IRV \leq .80$ were flagged as IER.

Psychometric Synonyms (*PS*) is a second measure of response consistency and refers to within-respondent correlations of vectors composed of items that are highly correlated with each other (Curran, 2016; Meade & Craig, 2012). Vectors are formed from pairs of items with high correlations for the entire sample (>.60 per

Curran, 2016). For example, the second item "I talk up this organization to my friends as a great organization to work for" and the sixth item "I am proud to tell others that I am part of this organization" were correlated $r = .80$ and placed in separate vectors. The lower the correlation between vectors, the less consistent respondents were across highly-related items. Curran recommended negative correlations as a conservative flag for IER. Individuals with negative correlations were coded "1" indicating IER; all others coded "0."

Effect Size Comparison. We calculated latent variable correlations, both with and without individuals identified as IER responders included, for each sample and compared those estimates to a meta-analytic referent. We used the uncorrected mean correlation for OC and WI of -.47 ($K = 351$, $N = 136{,}270$) from a meta-analysis by Cooper-Hakim and Viswesvaran (2005).

Results

Participant Demographics

Both MTurk (M_{Age}=34.1, SD_{Age}=10.6) and Qualtrics participants (M_{Age}=43.0, SD_{Age}=22.0) were, on average, older than student participants (M_{Age}=19.1, SD_{Age}=1.7). The percentage of females ranged from 51.3% (Qualtrics) to 54.9% (MTurk). Race and ethnicity were not available for students. Both MTurk and Qualtrics samples had demographic breakouts reasonably representative of the U.S. workforce (approximately 75% Caucasian, 8% African-American, and 5-10% Hispanic across samples). Hours worked per week varied between the Student and other two samples. Sixty-eight percent of students worked 20 hours per week or

less, while only 8.4% of the MTurkers and 4.7% of Qualtrics participants reported doing so.[2]

Replicating Population Effects

Pursuant to RQ1, latent variable correlations between OC and WI are presented in Table 1. Student and MTurk sample correlations were well above the -.47 meta-analytic correlation, even exceeding a .80 threshold for discriminant validity (Brown, 2006). The full Qualtrics sample (ρ = -.53) most closely approximated the meta-analytic referent.

Response Quality

Table 2 shows IER by metric for each sample as well as variability in the IER frequency across indices and samples. While the percentage of individuals flagged for IER was low and relatively consistent for *MD* and *PS*, the percent flagged was higher and more variable for *PTr* [ranging from 10.9% (Qualtrics) to 16.8% (MTurk)]. There were even greater inconsistencies for IRV values, ranging from 5.7% (MTurk) to 24.7% (Qualtrics).

Table 2 also shows the total number of respondents in each sample with zero, one, two, three, and four IER indicators. As can be seen, the Qualtrics sample had notably fewer respondents with no IER indicators (59.4%) than either student (71.2%) or MTurk (72.9%) samples. Respondents demonstrating IER generally were flagged by only one of four metrics.

[2] Detailed breakouts are available from the first author.

The lower half of Table 2 displays correlations between raw values for IER metrics by sample. Our two outlier indices (*PTr*, *MD)* are moderately and positively correlated across samples. Our two response consistency indices (*PS*, *IRV*) are positively correlated across samples and moderately correlated in two of three.

Impact of IER on Factor Correlations

Pursuant to RQ2, we examined factor correlations with and without IER respondents, and compared both to the meta-analytic correlation. We used latent variable modelling and conducted a multiple groups analysis with all parameters freely estimated.

Correlations for the full samples, and for sample subsets with IER respondents identified using each metric removed, are presented in Table 2. Correlations were, as expected, all negative, but most still stronger than our meta-analytic referent. Correlations were mainly similar with and without *MD* removed. This is not surprising, given the small percentage in each sample identified as IER by this metric. In contrast, correlations increased for each sample when those identified as IER by *IRV* were removed. The largest increase was for the Qualtrics sample (-.53 to -.78), perhaps because this sample had the largest proportion of IER removed (n = 79, 25%).

Correlations between the full samples and those with IER due to *PS* were similar, but correlations decreased (and were more similar to our meta-analytic referent) when IER respondents due to *PTr* were removed. To better understand the effect of removing IER due to *PTr*, we compared means on OC and WI between

participants coded as "1" and "0" in each sample. Means for each variable were significantly lower for OC and higher for WI in each sample for participants demonstrating IER. Visually, the impact of removing high IER participants by *PTr* was to eliminate participants tightly clustered in the upper left-hand corner of a scatter plot (extremely low on OC, extremely high on WI).

Discussion

Our study extends discussions on the representativeness and external validity of crowdsourced samples in multiple ways. These are summarized below.

Replicating the Population Effect Size

We examined the extent to which sample correlations replicated a meta-analytically-derived correlation between OC and WI. No samples generated a factor correlation within the 95% confidence intervals reported by Cooper-Hakim and Viswesvaran (-.49, -.45; 2005). It is worth noting that latent variable correlations are not the same as meta-analytically derived population estimates; whereas the former addresses measurement error, the latter also may address sampling error and range restriction.

Response Quality

We compared samples on four IER measures. Consistent with prior research (DeSimone & Harms, 2017; Meade & Craig, 2012), the percentage of IER responders depends on the metric. However, between 28% and 40% of respondents across samples demonstrated IER by at least one index. This is considerably higher than the 8-12% "modal rate" estimated by Curran (2016) and may be due to either

our use of multiple indicators or the specific indicators we chose. *IRV* values were the most inconsistent by sample type. Student and Qualtrics samples had the most IER respondents detected on *IRV*, MTurk on *PTr* values.

IER and Estimates of Factor Correlations

The Cooper-Hakim and Viswesvaran (2005) meta-analysis contained no studies removing IER participants. Thus, their meta-analysis may have mis-estimated effect size estimates in samples without IER. However, we know that removing of IER produces inconsistent effects on sample estimates across studies (compare Huang et al., 2015; McGrath et al. 2010). It is still premature to predict whether removing IER inflates or decreases correlations, but our results draw attention to the importance of being transparent about how IER is identified and treated. We suggest investigators routinely report findings with and without removal of IER - regardless of the sample.

Limitations

There are several limitations of our study in terms of how response quality was operationalized. First, we used no proactive or direct IER measure, which are common methods for detecting insufficient effort (e.g., Huang and colleagues, 2012, 2015). Using additional methods may have provided more insights into which participants were carelessly responding.

Second, because new methods for detecting IER are emerging, there are not yet clear benchmarks for labeling responses as IER. Other studies could use different cutoffs and reach different conclusions about the effects of IER; varying

cutoffs could result in different estimates of IER prevalence. We used our best judgment in identifying appropriate cutoffs by reading and following suggested guidelines, but it is clear further psychometric research is necessary to compare and validate cutoffs for identifying IER.

Theoretical Implications and Future Research

Our findings contribute to ongoing research and theory development on IER. Huang et al. (2012) defined IER as a response set in which survey items are answered with little desire to comply with survey instructions, interpret item content correctly, and/or respond accurately. Nichols, Greene, and Schmolck (1989) provided a useful distinction between response patterns: content responsive misrepresentation and content nonresponsive inaccuracy. The former suggests respondents understand what is being asked of them, but intentionally provide answers that misrepresent their "true scores" on constructs being measured. It is a form of impression management and includes well-known response behaviors like faking and social desirability.

In contrast, content nonresponsive inaccuracy occurs when participants do not read or interpret items accurately (Nichols et al., 1989). The key is that participant behavior is irrespective of item content. Nonresponsive inaccuracy can occur through the class of behaviors including IER. While IER and careless responding have been used interchangeably, we believe IER more accurately describes participant behavior. Respondents are often careful, but not in ways that shows sufficient effort responding to item content.

As more response quality metrics are proposed, tested, and routinely collected, it is possible researchers can build a nomological network relating metrics to respondent characteristics and motivation, as well as on their effects on study outcomes. As we demonstrated, IER metrics show some evidence of convergent validity, suggesting response consistency and outlier responses may be different constructs. Our analysis of factor correlations with and without IER showed that while generally removing IER respondents did not have a large effect on sample estimates, removing *IRV* tended to increase estimated factor correlations and removing *PTr* tended to decrease them. We suggest IRV assesses a form of systematic error that has the specific effect of lowering variance in observed variables; removing respondents exhibiting IER based on IRV should generally increase correlation estimates. In contrast, by identifying individual outliers, *PTr* may capture a form of systematic error and removing it should decrease observed correlations. Future researchers are encouraged to continue to estimate IER by multiple metrics and explore their effects on estimates of population parameters.

Practical Implications

Our results highlight factors to be considered when choosing a research sample. Both MTurk and Qualtrics samples were relatively representative of the workforce; hence organizational samples, the so-called "gold standard," are not inherently the most representative (Landers & Behrend, 2015). Sample representativeness and quality is also affected by the attentiveness of participant responses. Researchers are encouraged to continue using multiple samples to study

social phenomena, but to be careful to influence, assess, and control for participant IER.

Research on metrics to detect IER in surveys is relatively new and new measures and cutoffs are still emerging (e.g., Curran, 2016; Dunn et al., 2018; Huang & colleagues 2012, 2015; Meade & Craig, 2012). It is probable different IER indices tap different underlying constructs (cf., DeSimone & Harms, 2017). Researchers are encouraged to use multiple indicators of IER. Metrics may be chosen based on measure type, e.g., response consistency measures may be more diagnostic when scales contain reverse scored items.

We have several recommendations for survey research. First, we recommend researchers routinely and transparently present all information about their treatment of IER respondents. This includes which IER detection measures were used, what cutoffs were used to classify respondents as careless, and what proportion of the sample was so classified. Results with and without IER participants should be published or available on request. Second, as treatment of IER becomes more prevalent, we recommend in future meta-analyses researchers use IER treatment as a methodological moderator. Both type of treatment and percentage of participants excluded should be recorded so researchers can begin to better understand how IER treatment affects sample statistics.

Conclusion

Previously-held perceptions regarding the quality attributable to the data source need to be reconsidered. Our results revealed that while our Student sample

was not representative of the U.S. labor force, both MTurk and Qualtrics samples were reasonably so. Participants identified as careless responders depends greatly on the IER metric, but between 30 and 40% of participants in each sample demonstrated IER by at least one metric. When compared to a meta-analytic estimate of the relationship between OC and WI, removing careless respondents either increased or decreased the correlations depending on the metric.

References

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Cheung, J.H., Burns, D.K., Sinclair, R.R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology, 32*, 347-361.

Cooper-Hakim, A., & Viswesvaran, C. (2005). The construct of work commitment: Testing an integrative framework. *Psychological Bulletin, 131*, 241–259.

Crump, M.J., McDonnell, J.V., & Gureckis, T.M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8(3)*, e57410.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

Curran, P.G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4-19.

DeSimone, J.A., & Harms, P.D. (2017). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, 1-19.

Dunn, A.M., Heggestad, E.D., Shannock, L.R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology, 33*, 105-121.

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

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*, 99–114.

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

Landers, R.N., & Behrend, T.S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*, 142-164.

McGonagle, A.K., Huang, J.L., & Walsh, B.M. (2016). Insufficient effort survey

responding: An under-appreciated problem in work and organizational health psychology research. *Applied Psychology: An International Review, 65*, 287–321.

McGrath, R.E., Mitchell, M., Kim, B.H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*, 450-470.

Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.

Mobley, W.H., Horner, S.O., & Hollingsworth, A.T. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology*, *63*, 408-414.

Mowday, R.T., Steers, R.M., & Porter, L.W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, *14*, 224-247.

Mullinix, K.J., Leeper, T.J., Druckman, J.N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*, 109-138.

Nichols, D.S., Greene, R.L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*, 239-250.

Paolacci, G., & Chandler, J. (2014). Inside the Turk understanding Mechanical Turk as a student sample. *Current Directions in Psychological Science*, *23*, 184-188.

Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on Amazon

Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Walter, S.L., Seibert, S.E., Goering, D., & O'Boyle, E.H., Jr. (2018). A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology*.

Table 2. Factor Correlations With and Without IER

	All		Without...							
			MD		IRV		PS		PTr	
Sample	***N***	***r***	***N***	***r***	***N***	***r***	***N***	***r***	***N***	***r***
Student	382	-.80	380	-.80	331	-.84	352	-.80	337	-.71
MTurk	581	-.84	574	-.83	548	-.86	550	-.84	483	-.70
Qualtrics	320	-.53	313	-.53	241	-.78	295	-.51	285	-.36

Table 3. Count and Percentage of IER Analyses by Sample

IER Indicator	Student (N= 382)	MTurk (N= 581)	Qualtrics (N= 320)
Mahalanobis Distance	2 (0.5%)	7 (1.2%)	7 (2.2%)
Individual response variability	51 (13.4%)	33 (5.7%)	79 (24.7%)
Psychometrics Synonyms	30 (7.9%)	31 (5.3%)	25 (7.8%)
Person-Total correlations	45 (11.8%)	98 (16.8%)	35 (10.9%)
No IER Indicators	272 (71.2%)	424 (72.9%)	190 (59.4%)
1 Indicator	94 (24.6%)	148 (25.4%)	116 (36.3%)
2 Indicators	14 (3.7%)	7 (1.2%)	12 (3.8%)
3 Indicators	2 (0.5%)	3 (.3%)	2 (0.6%)
4 Indicators	0 (0.0%)	0 (0%)	0 (0%)
$r_{PS \cdot PT}$	-.71**	-.38**	-.64**
$r_{PS \cdot IVR}$	-.42**	-.40**	-.20**
$r_{PS \cdot MD}$	-.69**	-.47**	-.55**
$r_{PTr \cdot IVR}$	-.36**	-.02**	-.02**
$r_{PTr \cdot MDceu}$	-.58**	.30**	-.46**

$r_{IVR \cdot MD}$	-.09**	.20**	-.29**

**$p < .01$; *$p < .05$