

Convenience Samples and Measurement Equivalence in Replication Research

Lindsay J. Alley

Jordan Axt

Jessica Kay Flake

Psychology Department, McGill University

Contribution statement: LJA and JKF developed the idea and designed the study. LJA wrote the code and analysed the data. All authors contributed to writing the manuscript. JKF supervised the project.

Open Science statement: all materials, including code and data, available at <https://osf.io/ht48z/>

Abstract

A great deal of research in psychology employs either university student or online crowdsourced convenience samples (Chandler & Shapiro, 2016; Strickland & Stoops, 2019) and there is evidence that these groups differ in meaningful ways (Behrend et al., 2011). This could result in the presence of unaccounted-for measurement differences across convenience sample sources, which may bias results when these groups are compared, or the resulting data are pooled. In this registered report, we used the openly available data from the Many Labs replication projects to test for measurement equivalence across different convenience sample sources. We examined 9 measures that showed acceptable baseline model fit and tested them for non-equivalence across convenience samples from different sources, including university participant pools, MTurk, and Project Implicit. We then examined whether replication results are robust to non-equivalence by fitting partial invariance models and sensitivity analyses of replication results. [Results and discussion summarized here.]

Convenience Samples and Measurement Equivalence in Replication Research

In recent years, concerns about replication have become a source of interest and anxiety in many scientific fields, including psychology, genetics, cancer research, neuroscience, and economics (Zwaan et al., 2017). This is due, at least in part, to large collaborative projects that have attempted to estimate the rate at which findings replicate. One series of collaborations, called the Many Labs projects, has pooled resources across hundreds of scientists to collect large datasets for dozens of replication studies. There are five completed Many Labs studies (Ebersole et al., 2016, 2020; Klein et al., 2019, 2014, 2018), all involving large-scale collaboration of scientists and the pooling of data. Across all 62 effects replicated as part of these projects, 30 (48%) showed statistically significant effects in the same direction as the original study. Many scientists feel that the replication rates found by Many Labs and other similar projects are lower than they ought to be (Baker, 2016), and several statistical reforms meant to increase the replicability of the scientific literature have been discussed as a result (Shrout & Rodgers, 2018).

However, there has also been debate about the meaning of failed replications and what evidence they provide about the existence of any particular effect, as there are many features of both replications and original studies that could impact results. Various causes of failed replications have been discussed in the literature: lack of statistical power (Maxwell et al., 2015; Shrout & Rodgers, 2018), deviations from original methods in replication attempts (Gilbert et al., 2016), issues of research design and sampling (Nosek et al., 2022; Shrout & Rodgers, 2018), and measurement challenges (Fabrigar et al., 2020; Loken & Gelman, 2017). Though not often discussed, aspects of measurement can complicate the interpretation of replication results, including measurement differences between the original study and the replication, low reliability, lack of validity evidence, and measurement differences across relevant groups (Flake et al., 2022; Markon, n.d.; Shaw et al., 2020). Measurement

4 CONVENIENCE SAMPLES AND MEASUREMENT EQUIVALENCE

differences across groups often arise because people from varying backgrounds interpret items differently or use response scales in a dissimilar way. When this happens, the measure is said to be non-equivalent for those groups. The focus of this registered report is to consider the measurement equivalence (ME) of instruments collected as part of the Many Labs projects across two forms of convenience samples, specifically student and online crowdsourced samples. To introduce the study, we discuss measurement and replication, explain ME in more detail, and review the literature on measurement differences across convenience samples.

Measurement and Replication Research

In psychology, because the constructs we are interested in are not directly observable, researchers rely heavily on self-report scales, which aim to quantify unobservable psychological features, such as attitudes, moods, and personality traits. However, if researchers throw together a series of questions, they can't merely have faith that adding up the responses will result in a meaningful measure of the intended construct: they need to verify the validity and reliability of the scores they create or use (American Educational Research Association et al., 2014). Reviews of the psychological literature have found that the validity evidence presented by researchers does not live up to the standards of best scientific practice (Flake et al., 2017; Hogan & Agnello, 2004; Slaney, 2017). Reflecting the state of the field in general, the measures used in replication projects tend to have little validity evidence (Flake et al., 2022), and the Many Labs projects are no exception (Shaw et al., 2020). For instance, a review of all the measures used in Many Labs 2 (Shaw et al., 2020) found that 30% reported no reliability coefficients or validity evidence whatsoever and only 19% had a cited source. Additionally, Shaw et al. (2020) examined psychometrics of the measures using the open data from Many Labs 2 and found that most measures performed poorly according to common disciplinary standards: of the six scales examined, none met all

three fit index cut-offs selected (root mean square error of approximation [RMSEA] $< .05$, comparative fit index [CFI] $> .95$, standardized root mean squared residual [SRMR] $< .08$).

Large replication projects such as the Many Labs present a host of measurement challenges. The international and collaborative data collection is a strength (Henrich et al., 2010), but the pooling of data from heterogeneous samples can also introduce invalidity. When samples are drawn from different populations, there is the possibility that measures exhibit non-equivalence because the items do not hold the same meaning across populations. This poses a problem for replication projects, as ME is a prerequisite for valid group comparisons and the pooling of data across samples (Davidov et al., 2014).

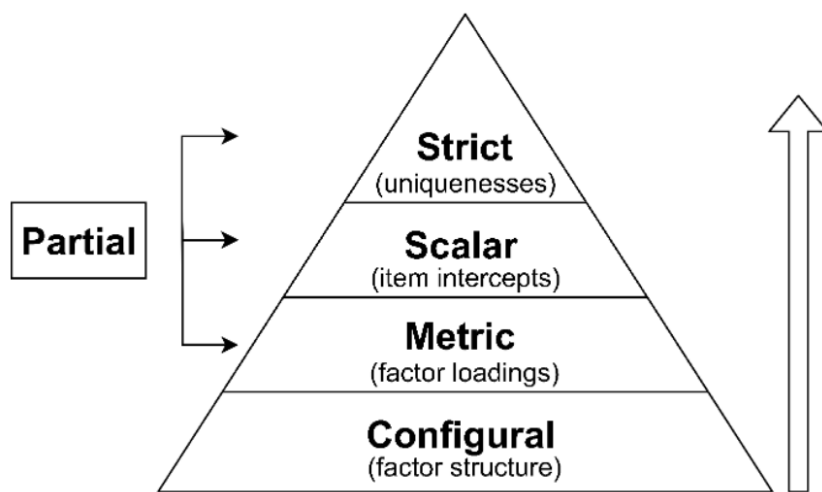
Two types of data sources are pooled in four of the five completed Many Labs projects: student samples and crowdsourced online samples. Because there are notable differences between these populations (Weigold & Weigold, 2021), there is a possibility this could introduce measurement non-equivalence, which might subsequently impact replication results. Though not a focus of the Many Labs projects at the outset, the open data and materials make it possible to evaluate ME after the fact. In this registered report, we propose to use a multiple group confirmatory factor analytic (MG-CFA) approach to test whether the measures employed in the Many Labs studies are equivalent across student samples and crowdsourced online samples, such as Amazon Mechanical Turk (MTurk). Confirmatory factor analysis (CFA) is a statistical modelling approach which aims to represent “the causal relations between one or more unobserved, or latent, variables and a set of observed variables” (Flora, 2017), and MG-CFA is the extension of this approach to model multi-group data, allowing for the detection and modelling of differences due to group membership. Next, we will complete a sensitivity analysis to understand if correcting for non-equivalence changes the results of the replication studies. Though the Many Labs projects are already completed, our results will help future researchers who hope to conduct large-scale collaborative research to understand whether variation across convenience samples is likely to be a meaningful and impactful source of

measurement non-equivalence, allowing researchers to account for this possibility in their analyses.

What is Measurement Equivalence?

Also called measurement invariance, measurement equivalence is concerned with whether a particular scale is measuring the same thing in the same way across different groups. Formally, this means that, for a given level of the latent trait, the conditional distribution of the items of the measure is the same across subpopulations (Meredith & Millsap, 1992). Thus, within a latent variable modelling framework, “measuring something in the same way” means that the items of the scale are related to the latent variable in the same manner across groups. There are different levels or degrees of ME, each of which has as its focus a different aspect of the item to latent variable relationship. These hierarchical, increasingly restrictive models can be tested using multiple group CFA, allowing researchers to understand to what degree the measures function in the same way across groups. Figure 1 shows an overview of the hierarchical levels of measurement equivalence; they are described in more detail below.

Figure 1.



Note: Overview of the Four Levels of Measurement Equivalence. Reprinted from “Measurement Invariance Testing Using Confirmatory Factor Analysis and Alignment Optimization,” by R. Luong and J. K. Flake, 2022, *Psychological Methods*, *Advance online publication*, p. 3. Copyright 2022 by the American Psychological Association.

The least restrictive level of ME is referred to as configural equivalence (Horn et al., 1983). This level requires that the number of latent factors, and which items load onto which factors, are the same across groups. In the case of scales intended to tap a single construct, this means that a unidimensional model must show adequate fit in both groups. The next level, commonly known as metric or weak equivalence, concerns the equivalence of the factor loadings across groups. The factor loadings represent the strength of the relationship between the individual items and the latent variable (Bollen, 1989); thus, metric equivalence is achieved when the slope of the item’s regression on the latent variable is the same across groups. The third level of equivalence is concerned with the intercepts of the item in the latent variable regressions and is called scalar or strong equivalence. Scalar non-equivalence occurs when one group uses the response scale for a particular item differently than another group, yielding mean items responses that are systematically higher or lower though their levels of the latent trait are the same (Cheung, 2008). Finally, the equivalence of error variances, or strict equivalence, should be considered. This will indicate whether items relate to the construct with the same degree of precision across groups.

When both metric and scalar equivalence are achieved, this is called strong factorial invariance. This is considered by many to be a prerequisite for using observed scores to make valid group comparisons (Cheung, 2008). Though MG-CFA can correct non-equivalence by estimating factor scores that take into account measurement differences, it isn’t standard practice for researchers in the social sciences: even among studies that compared across cultures, where non-equivalence is highly plausible, a review conducted by Boer et al. (2018)

found that only 13% of included studies tested for ME. Instead, researchers commonly calculate and compare sum scores (McNeish & Wolf, 2020). If there is non-equivalence across the groups measured and observed scores are used, these scores will be biased: intercept non-equivalence will bias group mean estimates and impact the results of t-tests (Steinmetz, 2013), while loading non-equivalence will impact regression coefficients and correlations (Chen, 2008).

Though ME is highly relevant to replication research, very little work has explored this intersection. As is the case with the Many Labs projects, many replications are conducted as part of large collaborative efforts where data from multiple populations are pooled. Even if replicators carry out the same research protocol and analyses, the conceptual interpretation of the items may be different across the different populations included in the study. If this is the case, the pooling of these data is not justified, and the presence of non-equivalence could bias results. Moreover, examination of the generalizability of the replication results across groups is compromised, as bias due to measurement non-equivalence may account for group differences regarding the effect of interest. In addition to being highly relevant to replication projects, these concerns apply to any “big team science” that pools data from many sources.

It would also be ideal for replication researchers to test for ME between original and replication studies (Fabrigar & Wegener, 2016), but this is difficult in practice: for the most part, the original studies that are replicated do not have publicly available data and have small sample sizes (Fraleay & Vazire, 2014). This is a barrier to detecting measurement non-equivalence, as sample sizes of approximately 400 per group are recommended to detect meaningful effects (French & Finch, 2016; Koziol & Bovaird, 2018; Meade & Bauer, 2007). However, large replication projects such as Many Labs make their data publicly available, enabling the assessment of ME across groups within the replications, such as data collection labs, translated versions of measures, and different sample sources. In this registered report, we made use of the

availability of these data to examine measure equivalence across student and crowdsourced convenience samples, two sample sources which are pooled in three of the five Many Labs projects.

Comparing Convenience Samples

University students and online crowdsourced samples are examples of different convenience samples. Baker et al. (2013) define convenience sampling as a non-probability data collection method that prioritizes “the ease with which potential participants can be located or recruited” (p. 94). The use of university student samples has been a popular form of convenience sampling in psychology for a long time, and the popularity of online crowdsourced samples is growing (Chandler & Shapiro, 2016; Strickland & Stoops, 2019). It is no wonder crowdsourcing research is becoming more popular: this approach offers many advantages, including cost-effectiveness, the ability to collect large samples quickly, and the potential to access diverse and hard to reach samples (Chandler & Shapiro, 2016; Strickland & Stoops, 2019). However, Strickland & Stoops (2019) point out that crowdsourced samples may differ from “the populations to which the results ideally would generalize” (p. 9), a type of selection bias. To deal with this limitation, they recommend that researchers collect samples through diverse methods and consider aggregate results. If this approach is to be effective, it’s important that aggregated samples demonstrate ME, or that researchers employ a statistical model that accounts for non-equivalence across samples. If selection bias and hidden measurement differences are both impacting the results of a study, it is important to correct for ME in order to disentangle these two sources of bias.

MTurk is one of the most popular platforms for crowdsourcing research participants, due to its large user base, affordability, and ease of use. As such, a great deal of the research comparing crowdsourced and student convenience samples focuses specifically on MTurk and has found that student and MTurk samples tend to differ in several ways.

Demographically, MTurk samples are consistently older than student samples (Behrend et al., 2011; Roulin, 2015; Steelman et al., 2014), often more ethnically diverse (Behrend et al., 2011), and come from a lower socioeconomic background (Weigold & Weigold, 2021). Additionally, though college students can be recruited through MTurk, they tend to be farther along in their degrees and are more likely to be part time compared to those recruited through university participant pools (Weigold & Weigold, 2021). MTurk and student samples also show mean differences on measures of personality: student samples are reliably higher in extraversion (Behrend et al., 2011; Colman et al., 2018; Goodman et al., 2013; Weigold & Weigold, 2021), and MTurk samples tend to score higher on openness to experience (Behrend et al., 2011; Colman et al., 2018; Weigold & Weigold, 2021). MTurk is the most studied online crowdsourcing platform, but research on differences from student samples may not generalize to other, similar data-collection platforms. Peer et al. (2022) found that data from Prolific, CloudResearch, Qualtrics, and Dynata differed from MTurk in terms of demographics and data quality. While differences across samples do not necessarily indicate non-equivalence, differences in sample characteristics could potentially contribute to non-equivalent measurement for particular constructs, as respondents from groups that differ from each other may understand items differently. However, it is also possible that very different people interpret items in the same way, and, therefore, these groups could still be equivalent in terms of measurement properties for a given construct. It is important to examine the issue directly.

There is a small but growing body of research on ME between student and MTurk samples. One study investigating a measure of post-traumatic stress disorder symptomology concluded that strict ME held across these samples (Caldas et al., 2020). Other studies that found equivalence across these samples, examining a multi-faceted personality disorder measure and measures of openness and innovation respectively, only examined configural (Miller et al., 2017) or loading equivalence (Winton & Sabol, 2021), leaving the equivalence of intercepts and error variances untested. Additionally, Behrend et al. (2011) assessed the

equivalence of measures of Big Five personality traits and goal orientation and found that, while a few items from these scales were non-equivalent across groups, the effect sizes were small enough that the scales were functionally equivalent.

Adding some complexity to the issue, no two MTurk samples are the same and can vary in terms of culture and English language ability, as MTurkers can be recruited from all over the world. For instance, Feitosa et al. (2015) found that a measure of Big Five personality traits was equivalent to the scalar level between a student and a US-only MTurk sample, but only configural equivalence held when students were compared with a non-US MTurk sample. As this non-US sample was composed largely of non-native English speakers from India, they conclude that equivalence may not hold when MTurkers first language is not English.

In this registered report, we will extend and build on previous work in three important ways. First, we will conduct a thorough investigation of ME for a set of untested scales. While previous work has tested the equivalence of a number of measures, this does not mean that the same conclusions will be reached for different measures. Equivalence is sensitive to the construct being measured and the specific wording of items, so what holds for one measure may not for others. Second, we will be able to examine a source of crowdsourced data other than MTurk, as Many Labs 1 also includes a sample collected through Project Implicit (see Table 1 for a breakdown of sample sizes by source). This extends the literature on this topic because MTurk samples are used almost exclusively to represent all online crowdsourced samples, but there is no guarantee that the results would generalize to other similar sources. Third, Many Labs 2 includes an MTurk sample from India and one from the US, which will allow us to test whether prior work on the importance of language spoken (Feitosa et al., 2015) is found in a new set of measures. Overall, this study is the most

comprehensive examination of ME between convenience samples to date, in terms of the number of measures examined, the variety of sample sources, and sample size.

ML1	MTurk: 1000 Implicit: 1329 Student (lab): 2404 Student (online): 737
ML2 (slate 1)	MTurk (India): 360 MTurk (US): 331 Student (lab): 2557 Student (online): 256
ML2 (slate 2)	MTurk (India): 362 MTurk (US): 340 Student (lab): 1885 Student (online): 1467
ML3	MTurk: 737 Student: 2741

Table 1. Sample sources in each Many Labs project and total sample size per source.

Our analyses are driven by two primary research questions:

RQ1. To what extent do measures function equivalently across different convenience samples in the Many Labs projects?

RQ2. When measures are non-equivalent, does correcting for this change the statistical significance or effect sizes of the replications?

Answering these questions will contribute to understanding and addressing methodological challenges that are present in replication projects and beyond. First, previous research has not explored the degree to which a lack of ME across samples in replications and other collaborative projects presents an issue, both in terms of prevalence (RQ1) and impact (RQ2). By examining the issue for convenience samples, we can begin to explore the scope of this problem for one possible source of non-equivalence. Second, to the extent that measurement non-equivalence presents a problem, the analyses that we present here may serve as a template for researchers to consider ME as a part of their analysis plan in future replications and collaborative research projects and, based on our experience completing

these analyses, we can make recommendations that may contribute to best practices moving forward. Finally, the results of this project will contribute to understanding whether different convenience sample sources tend to display measurement non-equivalence by examining multiple measures, which is useful more broadly than just replication research, especially given how common these sample sources are in psychology. Understanding whether different convenience samples are likely to display measurement non-equivalence will aid in the interpretation of all studies that use these samples and contribute to building a cumulative psychological science. For an overview of the design of our study to answer each of our research questions, see the Study Design Table in our supplementary materials.

Methods

In the following section, we describe in detail the preliminary measure inclusion analyses and the analyses for the main questions of interest, the equivalence testing and sensitivity analysis. [Note: measure inclusion analyses were performed before the submission of the stage 1 manuscript. The other analyses have not yet been completed.] Code for all analyses can be found in the supplementary materials.

Preliminary Measure Inclusion Analyses

The primary proposed analysis is psychometric equivalence testing. We performed these tests using MG-CFA with maximum likelihood estimation, which requires that the data meet the assumptions of the estimation method (multivariate normal, sufficient response options to approximate continuous) and, additionally, that the baseline measurement model is adequately specified (French & Finch, 2011). To determine which scales are amenable to the analyses, we carried out Confirmatory Factor Analyses (CFAs) for all measures that met the following criteria: 4 or more items per factor, enough response options that the items may be treated as continuous (Rhemtulla et al., 2012), and completed by both student and online crowdsourced samples (see Table 2 for scale information and CFA results). Type I error rates for equivalence tests may be inflated when the baseline model is misspecified (French & Finch,

2011), resulting in a higher probability of incorrectly concluding that a scale is non-equivalent across groups. For example, if a measure is modelled as unidimensional, but the items in fact load onto two factors, an equivalence test for this incorrectly specified unidimensional model would be more likely to find non-equivalence across groups, even though the true, 2-factor model is equivalent. For this study, we must balance the importance of controlling Type I error rates with the importance of investigating as wide of a range of instruments as possible. Given those considerations, we selected fit index cut-offs consistent with mediocre, but not clearly terrible fit: RMSEA $\leq .10$ (Browne & Cudeck, 1992), SRMR $\leq .10$ (Kline, 2015), CFI $\geq .90$ (Kline, 2015). We excluded models from further analyses which failed to meet two out of three of these cut-offs. Code for these analyses can be found in the supplementary materials (Inclusion Code).

Overall, five measures were eliminated, and nine remained as candidates for equivalence testing (see Table 2). These measures represent a diverse set of constructs, which can increase the generalizability of our conclusions. The nine measures selected for further analyses are briefly described below.

1. Contact Intentions (ML1 Study 11): this 4-item measure of respondents' future intentions to interact with Muslims was adapted by Husnu & Crisp (2010) from a measure of behavioural intentions (Ratcliff et al., 1999). Replicators changed the items to refer to Muslims more generally rather than British Muslims, as in the original study.
2. Explicit Math Attitudes (ML1 Study 13): measures the valence of respondents' attitudes towards math using six Likert items and one 100-point feelings thermometer. This measure was developed by authors for a study of explicit and implicit attitudes towards math across genders (Nosek et al., 2002), replicators used a subset of items.
3. & 4. Moral Foundations Questionnaire - Individualizing and Binding (ML2 Study 4): developed by Graham et al. (2009) to measure the relevance to moral decision-making of

their theorized five moral foundations: harm, fairness, ingroup, authority, and purity.

These foundations were assessed using 15 Likert items, three per foundation, which were further grouped into the higher-order factors of individualizing and binding moral foundations: the harm and fairness foundations are grouped under individualizing, and the ingroup, authority, and purity foundations form the binding factor. For the replication, this measure was scored by averaging responses to the items that form the higher-order individualizing and binding factors; for this reason, we examined Individualizing and Binding as separate scales.

5. Leader Power Scale (ML2 Study 15): a scale for rating the perceived power of a leader or manager, created by Giessner and Schubert (2007). This measure consists of five Likert-type items that assess the perceived dominance, confidence, and level of control that the target leader displays.

6. Desire for Control Products (ML2 Study 23): two scales were developed for use in a study by Zhong & Liljenquist (2006), one where respondents rated their desire for five different cleaning products, and this scale, where respondents rated their desire for an assortment of five other products (“control products”). While we considered both scales for inclusion in this study, only the Desire for Control Products scale met our fit criteria.

7. Argument Quality (ML3 Study 8): this five item scale, created by Cacioppo et al. (1983) for use in their study, asks respondents to rate the quality of a target piece of argumentative writing.

8. Need for Cognition (ML3 Study 8/Individual difference measure 5): the original study (Cacioppo et al., 1983) employed a 34-item measure of the need for cognition construct. According to the developers, this scale examines “the tendency for an individual to engage in and enjoy thinking” (Cacioppo & Petty, 1982, p. 116). Replicators used a

shortened version, consisting of the six items with the highest factor loadings in the validation literature.

9. Perceived Stress Scale (ML3 Individual difference measure 4): this scale was not part of any replicated effect but was employed to measure respondents' perceptions of their stress over the past week. A short-form scale consisting of four items was used (Cohen et al., 1983).

Scale	Items	Type	α	χ^2	df	CFI	RMSEA - 90%CI	SRMR
Political Attitudes (PA)	8	7-Point	.68	1251.94	20	0.80	0.11 [0.10, 0.11]	0.06*
System Justification (SJ)	8	7-Point	.78	1414.82	20	0.86	0.12 [0.11, 0.12]	0.06*
Contact Intentions (CI)	4	9-Point	.83	198.01	2	0.98*	0.14 [0.12, 0.15]	0.02*
Explicit Math Attitudes (EMA)	7	Mixed	.95	1034.01	14	0.97*	0.13 [0.12, 0.14]	0.02*
Moral Foundations Questionnaire Individualizing (MFQ-I)	6	6-Point	.82	271.12	9	0.97*	0.08* [0.07, 0.09]	0.03*
Moral Foundations Questionnaire Binding (MFQ-B)	9	6-Point	.78	1333.97	27	0.88	0.09* [0.09, 0.10]	0.05*
Subjective Well Being (SWB)	25	Mixed	.79	18653.19	275	0.41	0.16 [0.15, 0.16]	0.20
Leader Power (LP)	5	7-Point	.86	785.03	5	0.92*	0.19 [0.18, 0.20]	0.04*
Desire for Cleaning Products (D-Clean)	5	7-Point	.77	863.04	5	0.89	0.17 [0.16, 0.18]	0.06*
Desire for Control Products (D-Cont)	5	7-Point	.49	249.94	5	0.87	0.09* [0.08, 0.09]	0.04*
Argument Quality (AQ)	5	9-Point	.87	57.3	5	0.99*	0.08* [0.06, 0.09]	0.02*
Need for Cognition (NfC)	6	5-Point	.67	99.78	9	0.95*	0.06* [0.05, 0.07]	0.03*
Perceived Stress Scale (PSS)	4	5-Point	.72	93.03	2	0.96*	0.13 [0.11, 0.15]	0.03*
Intrinsic Motivation (IM)	15	4-Point	.79	5816.42	90	0.57	0.14 [0.14, 0.15]	0.12

Table 2. CFA results for all suitable measures, using total sample collected for each measure. * fit index meets proposed cut-off. Scales that qualify for further analyses are bold.

Analysis Plan

Code for the following analyses can be found in the supplementary materials. There is a separate R file for each measure, and the files are named for the measure analyzed (i.e. Contact Intentions Analyses, Explicit Math Attitudes Analyses etc.). The code used to develop the analysis plan can also be found in the supplementary materials (Planned Analysis

Code). Sections of code pertaining to the analyses described below are cited as (code x.x), and the sections are numbered in the same way for all code files. [Note: only the Planned Analysis Code is included at stage 1. The other files described will be added when full analyses are completed for stage 2.]

Demographics

We examined the available demographic variables by sample group for each Many Labs project included in this paper (1, 2, and 3) in the appropriate way for each variable type (mean and standard deviation for continuous variables, percentages for categorical variables like gender, code 1.2). There is some variation as to which variables were collected for each project: Many Labs 2 reports only age and gender, while Many Labs 1 and 3 collected a number of other demographic variables, such as ethnicity and native language.

Assumptions and Data Checks

To minimize the impact of assumption violations such as the lack of multivariate normality or model misspecification, we employed maximum likelihood estimation with Huber-White robust standard errors (MLR estimator in lavaan). However, we still examined some item level information to check that our data were reasonable after processing. Specifically, we examined skew, kurtosis, and item histograms and correlation matrices (code 1.3). Additionally, we fit single sample CFAs in the full data, and separately in each group, and examined the fit statistics and reliability (code 1.4).

Measurement Equivalence Analyses

In order to avoid conflating the issue of non-equivalence due to instrument translation with non-equivalence due to sample source, we limited our analyses to participants who completed the studies in English (code 1.1). The analyses were completed for all measures that fit the selection criteria, 9 scales in total. For each measure, each sample group was compared separately to each other sample group available for that measure. For example, in

Many Labs 1 there are four sample groups of interest, so equivalence was tested across six pairs of convenience sample types for every measure from that project: MTurk vs. Project Implicit, MTurk vs. student (lab), MTurk vs. student (online), Project Implicit vs. student (lab), Project Implicit vs. student (online), and student (lab) vs. student (online).

For ME testing, we used a hierarchical approach: we compared multiple group CFA models of increasing restrictiveness (equal factor structure, loadings, intercepts, residuals) and stopped when the additional restrictions were rejected (Byrne & van de Vijver, 2010; Luong & Flake, 2022) (code 2.2, 2.5). To set the scale of the latent variable, we fixed its mean to 0 and variance to 1 for one group and freely estimated these values for the other. To identify the model, it is also necessary to select an anchor item. This is an item which is presumed to be equal psychometrically across groups. By constraining the loading and intercept of this item to be equivalent across groups, this ensures that the scale of the latent variable is the same, which allows for the equivalence of other items to be tested. To determine the anchor item, we employed Likelihood ratio tests using the all-other-items-as-anchors approach (Woods, 2009): starting from a model with all loadings and intercepts constrained to be equal across groups, then freeing both parameters for one item at a time and comparing this to the constrained model. For each measure, the item with the smallest Likelihood ratio associated with this test was selected as the anchor item (code 2.1).

Many of the convenience sample groups we examined are of very different sizes, which can bias equivalence testing such that non-equivalence is more difficult to detect (Yoon & Lai, 2018). For any sample pairing which was substantially unbalanced (one sample 1.5 or more times the size of the other), we employed the subsampling method proposed by Yoon & Lai (2018) to force balance to the samples (code 2.2, 2.5).

In addition to unbalanced sample sizes, it is important to consider the impact of sample size on power, as results of statistical tests should be interpreted with caution in situations

where the power to detect a meaningful effect is insufficient. Power for the χ^2 -difference test of the equivalence of loadings and intercepts across groups is complex, as it is influenced not only by sample size and the amount and degree of non-equivalence, but also by many other features of the data and model, including: the strength of the loadings for non-equivalent items (Meade & Bauer, 2007), whether the direction of the non-equivalence is uniform or mixed (i.e. some loadings higher and some lower in the focal group, versus all loadings lower in the focal group; Meade & Bauer, 2007), the number of factors (French & Finch, 2006; Meade & Bauer, 2007), and the number of items per factor (Finch & French, 2018; French & Finch, 2006).

Simulation research on the χ^2 -difference test of the equivalence of loadings has found that, for sample sizes of 150 to 200 per group, power varies substantially based on these features (as low as .29 or as high as .95; French & Finch, 2016, 2006; Koziol & Bovaird, 2018; Meade & Bauer, 2007). For sample sizes of 400 to 500 per group, power is generally high: while one study reported power of .57 in a condition with 500 per group (French & Finch, 2006), this was an anomaly, and every other study reported values of .89 or greater (French & Finch, 2016; Koziol & Bovaird, 2018; Meade & Bauer, 2007). Of the 14 sample groups that we plan to examine, 5 of them have a sample size less than 400, and one of these is below 300 (the online student sample in ML2 slate 1). As such, we expect that results involving these sample groups should be interpreted with caution.

To evaluate the tenability of each level of parameter restrictions, we compared each nested model to the next most restricted one using Satorra and Bentler's (2001) approach to calculating the scaled χ^2 -difference statistic. A non-significant χ^2 -difference test indicates that the addition of the restricted parameters does not add an unacceptable degree of misfit and it is plausible that the relevant parameters are equal across groups in the population. If one of the χ^2 -difference tests was significant at $\alpha = .05$, this was taken to indicate non-equivalence at

that level (code 2.2, 2.5). Due to the fact that we may find statistically significant, but not practically significant non-equivalence, we also report dMACS effect sizes (Nye & Drasgow, 2011), though these were not used for decision making (code 2.3). Based on simulation studies by Nye et al. (2019), when less than 50% of the items are non-equivalent, we consider dMACS > .40 to be practically significant; and when 50% or more are non-equivalent, we consider dMACS > .20 to be practically significant.

If a particular measure was not equivalent between groups to the strict level, we stopped the hierarchical testing procedure at whichever level the additional restrictions were rejected and proceeded to test the equivalence of the items so that we could develop a partial equivalence model. This is necessary in order to complete the sensitivity analysis (RQ2) comparing results using sum scores to factor scores produce by the partial equivalence models. In order to identify which item parameters were non-equivalent, we employed univariate score tests (Bentler & Chou, 1992), also referred to as modification indices (code 2.4). We assessed the parameters iteratively, releasing the one with the largest χ^2 value and then testing the items again to identify any additional non-equivalent parameters. We proceeded until all score tests were non-significant, or the relevant parameter was only constrained for two items in the final model (Byrne et al., 1989). We used a Bonferroni corrected alpha level of .05 divided by the number of parameters being tested in that block. For example, if testing the loadings of an 8-item measure, the critical α would be .007, or .05 divided by one less than the total number of items, due to the anchor item remaining fixed (code 2.4). We only completed this process for loadings and intercepts; if strict equivalence was rejected, we allowed all error variances to differ across groups.

Sensitivity Analysis of Replication Effects

To examine the impact of measurement non-equivalence on the replicated effects, we reproduced the analyses conducted in the Many Labs for any measure that displayed some

level of non-equivalence across groups and is involved in a replication effect. We produced factor scores using the final partial equivalence MG-CFA model for that measure (code 3.1) and, using the openly available analysis code for each study, reproduced the replication analyses using these factor scores in the place of the sum or mean scores originally used (code 3.2). Because factor scores also correct for measurement error, using them could change the results of some analyses even in the absence of measurement non-equivalence. To isolate the specific effect of non-equivalence, we also reproduced the analyses using factor scores from single group CFAs (code 3.1, 3.2). Regression factor scores were used (Thurstone, 1935) because they exhibit less bias in the estimation of downstream effects compared to Bartlett's factor scores (Devlieger, Mayer, and Rosseel, 2016), the other factor score estimation method implemented in lavaan for continuous data.

Level of Bias Control

We submitted this registered report as designated at Level 2 bias control. This is because the data were already available at the time of analysis planning, and we had accessed the data to perform other analyses but had not separated the data by convenience sample source or performed any of the ME analyses for these groups. To further control for the risk of bias, we developed a detailed analysis plan including code. The Planned Analysis Code contains all proposed analyses completed using the real data for one measure, the 8-item Political Attitudes (PA) measure from Many Labs 1. However, we created a fake, randomly generated grouping variable rather than separating the data by sample source (Planned Analysis Code 1.1) to reduce the risk that we would make choices in order to achieve interesting results in the planning stage. We chose this measure for the purpose of analysis planning because it was eliminated from inclusion in the final study due to poor model fit, so we did not need to interact further with the portions of the data that would be used for our primary analyses.

Results

[Example results tables are included below with the results from our analysis planning.

Additionally, we will describe our results in text.]

Demographics

Example Data	Age	Sex	Native Language
MTurk	Range: 14-79 Mean: 27.6 (12.6)	Female: 543 (68%) Male: 255 (32%) Not reported: 2 (<1%)	English: 682 (85%) Spanish: 28 (4%) Other: 87 (11%)
Student	Range: 13-85 Mean: 26.4 (11.9)	Female: 3063 (67%) Male: 1486 (33%) Not reported: 8 (<1%)	English: 3950 (87%) Spanish: 135 (3%) Other: 451 (10%)
Many Labs 1	Age	Sex	Native Language
Project Implicit			
MTurk			
Student (lab)			
Student (online)			
Many Labs 2 (slate 1)	Age	Sex	
MTurk (India)			
MTurk (US)			
Student (lab)			
Student (online)			
Many Labs 2 (slate 2)	Age	Sex	
MTurk (India)			
MTurk (US)			
Student (lab)			
Student (online)			
Many Labs 3	Age	Sex	
MTurk			
Student			

Table 3. Demographics by sample group.

Measurement Equivalence

[illegible]

Table 4. Measurement equivalence test results: highest level of equivalence achieved for each sample is marked. Full statistical results available in supplementary materials: <https://osf.io/ht48z/>

24 CONVENIENCE SAMPLES AND MEASUREMENT EQUIVALENCE

Political Attitudes [Example for analysis planning, will be removed]	Item	1	2	3	4	5	6	7	8
	DMACS	0.06	0.09	0.08	0.05	0.13	0.13	0.08	0
Contact Intentions	Item								
	DMACS								
Explicit Math Attitudes	Item								
	DMACS								
Moral Foundations Questionnaire Individualizing	Item								
	DMACS								
Moral Foundations Questionnaire Binding	Item								
	DMACS								
Leader Power	Item								
	DMACS								
Desire for Control Products	Item								
	DMACS								
Argument Quality	Item								
	DMACS								
Need for Cognition	Item								
	DMACS								
Perceived Stress Scale	Item								
	DMACS								

Table 5. DMACS effect sizes. Suggested cut-offs for interpretation: >.20 and <.40 small, >.40 and <.70 medium, >.70 large (Nye et al., 2019). Anchor item, italicized, will always have a DMACS of 0. [Each section of this table will be adjusted to display the correct number of items for the relevant measure, as each measure is different.]

Scale	Partial Equivalence Model		
	Loadings freed	Intercepts freed	Error variances freed
Political Attitudes (MTurk vs Student) [Example for analysis planning, will be removed]	Item 2	None	None
Contact Intentions			
Explicit Math Attitudes			
Moral Foundations Questionnaire Individualizing			
Moral Foundations Questionnaire Binding			
Leader Power			
Desire for Control Products			
Argument Quality			
Need for Cognition			
Perceived Stress Scale			

Table 6. Descriptions of the partial equivalence models. [Note: all measures are listed in this example table, but only some measure/sample combinations will have partial equivalence models. If equivalence holds to at least scalar, no partial equivalence model will be developed and that measure will not be described here.]

Sensitivity Analysis

[In addition to describing the results, we will develop a plot to display effect sizes using factor scores vs original scoring method for all measures included in the sensitivity analysis.]

Discussion

- Results have implications for replications and other large scale collaborative projects that pool student/online sources:
 - If equivalence holds, this is justification for pooling these samples.
 - If some or all are non-equivalent, this emphasizes the importance of testing ME for these projects.
 - If there is undetected non-equivalence in replication studies, this may impact the meaning of the results. Given our findings, we will discuss the degree to which this may be a concern for convenience samples.
- This is also relevant to research more broadly: if the measurement of constructs is often functioning differently across these two very common sample types, then results may not be comparable in the way that they are commonly interpreted. Equivalence across these samples works to build a cumulative science beyond replication research.
- The results of our sensitivity analysis will speak to the degree to which any non-equivalence detected is of practical importance to researchers. We will discuss the results here, including whether and how non-equivalence contributed to the replication effects, and what features of the non-equivalence (loading vs intercept, effect size, number of items, direction of non-equivalence) impacted results.
- Based on these results, we will make some recommendations for when researchers should examine equivalence across these samples, how to incorporate these tests into their analyses, and what to do when samples are non-equivalent.
- Limitations:
 - While we are testing a large number of scales, there still exists a wide variety of other scales in use for which different results might be obtained. We can only contribute to understanding the overall trend of whether different

convenience sample sources are likely to contribute to non-equivalent measurement and cannot settle the question.

- Some of the baseline models may be misspecified, which can result in incorrect estimates of other parameters. Might impact ME test results.
 - Power: tests involving the 5 smaller samples may have had low power to detect meaningful non-equivalence. We will highlight which effects these are and discuss implications a lack of power might have had.
 - Lack of validity evidence for measures from Many Labs: Measurement equivalence is a test of whether a scale is measuring a construct in the same way across groups. If the scale is not, in fact, measuring any construct at all, this question ceases to make any sense.
- In order to examine equivalence across convenience samples in this project, we had to make decisions about how to deal with other plausible sources of non-equivalence. We opted to collapse across many groups, such as experimental conditions, participant gender, and participant race, all of which can contribute to non-equivalence. We also completely eliminated translated instruments, which are known to be a source of non-equivalence, by only using English versions of measures. If we considered every possible subgrouping, and clustered respondents into only those exactly like them, the groups would be too multitudinous and fine-grained to proceed with any examination. If we had sufficient data to do so, we could consider more groups simultaneously using the alignment method (Asparouhov & Muthén, 2014) for equivalence testing. However, given the available subgroup sample sizes in many cases, the issue necessitates some simplifying decisions regarding which features are likely to be relevant for a given measure. As a result, a limitation of this work is that we cannot be sure that the decisions we made are the right ones.

- Future studies could collect data from these convenience sample sources for a range of scales with strong previous validity evidence to conduct an even more thorough examination of potential measurement differences between these sample sources.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 495–508.
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454.
- Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology*, 1(2), 90–143.
- Bartlett, M. S. (1937). The Statistical Conception of Mental Factors. *British Journal of Psychology. General Section; London, Etc.*, 28(1), 97.
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3), 800–813.
- Bentler, P. M., & Chou, C.-P. (1992). Some New Covariance Structure Model Improvement Statistics. In *Sociological Methods & Research* (Vol. 21, Issue 2, pp. 259–282). <https://doi.org/10.1177/0049124192021002006>
- Boer, D., Hanke, K., & He, J. (2018). On Detecting Systematic Measurement Error in Cross-Cultural Research: A Review and Critical Reflection on Equivalence and Invariance Tests. *Journal of Cross-Cultural Psychology*, 49(5), 713–734.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.
- Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. In *Sociological Methods & Research* (Vol. 21, Issue 2, pp. 230–258). <https://doi.org/10.1177/0049124192021002005>
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor

covariance and mean structures: The issue of partial measurement invariance.

Psychological Bulletin, 105(3), 456–466.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for Measurement and Structural

Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of

Nonequivalence. *International Journal of Testing*, 10(2), 107–132.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and*

Social Psychology, 42(1), 116.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message

evaluation, recall, and persuasion. In *Journal of Personality and Social Psychology*

(Vol. 45, Issue 4, pp. 805–818). <https://doi.org/10.1037/0022-3514.45.4.805>

Caldas, S. V., Contractor, A. A., Koh, S., & Wang, L. (2020). Factor Structure and Multi-

Group Measurement Invariance of Posttraumatic Stress Disorder Symptoms Assessed

by the PCL-5. *Journal of Psychopathology and Behavioral Assessment*, 42(2), 364–

376.

Chandler, J., & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced

Convenience Samples. *Annual Review of Clinical Psychology*, 12, 53–81.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of

making inappropriate comparisons in cross-cultural research. *Journal of Personality*

and Social Psychology, 95(5), 1005–1018.

Cheung, G. W. (2008). Testing Equivalence in the Structure, Means, and Variances of

Higher-Order Constructs With Structural Equation Modeling. *Organizational*

Research Methods, 11(3), 593–613.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress.

Journal of Health and Social Behavior, 24(4), 385–396.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). *Measurement*

Equivalence in Cross-National Research. <https://doi.org/10.1146/annurev-soc-071913-043137>

- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82.
- Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., ... Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in Methods and Practices in Psychological Science*, 3(3), 309–331.
- Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, 66, 68–80.
- Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, 24(4), 316–344.
- Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, 75, 47–52.
- Finch, W. H., & French, B. F. (2018). A Simulation Investigation of the Performance of Invariance Assessment Using Equivalence Testing Procedures. *Structural Equation*

Modeling: A Multidisciplinary Journal, 25(5), 673–686.

- Flake, Jessica K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, 8(4), 370–378.
- Flake, Jessica Kay, Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *The American Psychologist*.
<https://doi.org/10.1037/amp0001006>
- Flora, D. B. (2017). *Statistical Methods for the Social and Behavioural Sciences: A Model-Based Approach*. SAGE.
- Fraley, R. C., & Vazire, S. (2014). The N-pact factor: evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, 9(10), e109019.
- French, B. F., & Finch, H. (2016). Factorial invariance testing under different levels of partial loading invariance within a multiple group confirmatory factor analysis model. *Journal of Modern Applied Statistical Methods: JMASM*, 15(1), 511–538.
- French, B. F., & Finch, W. H. (2006). Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3), 378–402.
- French, B. F., & Finch, W. H. (2011). Model Misspecification and Invariance Testing Using Confirmatory Factor Analytic Procedures. *Journal of Experimental Education*, 79(4), 404–428.
- Giessner, S. R., & Schubert, T. W. (2007). High in the hierarchy: How vertical location and judgments of leaders' power are interrelated. *Organizational Behavior and Human Decision Processes*, 104(1), 30–44.
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science.” In *Science* (Vol. 351, Issue 6277, pp. 1037–

1037). <https://doi.org/10.1126/science.aad7243>

- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 1029–1046.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29–29.
- Hogan, T. P., & Agnello, J. (2004). An Empirical Study of Reporting Practices Concerning Measurement Validity. *Educational and Psychological Measurement*, 64(5), 802–812.
- Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*, 1(4), 179–188.
- Husnu, S., & Crisp, R. J. (2010). Elaboration enhances the imagined contact effect. In *Journal of Experimental Social Psychology* (Vol. 46, Issue 6, pp. 943–950). <https://doi.org/10.1016/j.jesp.2010.05.014>
- Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Hilgard, J., Ahn, P. H., Brady, A. J., Chartier, C. R., Christopherson, C. D., & al., E. (2019). *Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement*. <https://doi.org/10.31234/osf.io/vef2c>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr, Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., ... Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3), 142–152.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard,

- M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzaska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490.
- Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.
- Kozioł, N. A., & Bovaird, J. A. (2018). The Impact of Model Parameterization and Estimation Methods on Tests of Measurement Invariance With Ordered Polytomous Data. *Educational and Psychological Measurement*, 78(2), 272–296.
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584–585.
- Luong, R., & Flake, J. K. (2022). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*. <https://doi.org/10.1037/met0000441>
- Markon, K. E. (n.d.). *Reliability, Replicability, and Validity: A Meta-Scientific Analysis*.
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *The American Psychologist*, 70(6), 487.
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52(6), 2287–2305.
- Meade, A. W., & Bauer, D. J. (2007). Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 611–635.
- Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection

of measurement bias. In *Psychometrika* (Vol. 57, Issue 2, pp. 289–311).

<https://doi.org/10.1007/bf02294510>

- Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personality Disorders*, 8(1), 26–34.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math= male, me= female, therefore math≠ me. *Journal of Personality and Social Psychology*, 83(1), 44.
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73, 719–748.
- Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How Big Are My Effects? Examining the Magnitude of Effect Sizes in Studies of Measurement Equivalence. *Organizational Research Methods*, 22(3), 678–709.
- Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: understanding the practical importance of differences between groups. *The Journal of Applied Psychology*, 96(5), 966–980.
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643–1662.
- Ratcliff, C. D., Czuchry, M., Scarberry, N. C., Thomas, J. C., Dansereau, D. F., & Lord, C. G. (1999). Effects of directed thinking on intentions to engage in beneficial activities: Actions versus Reasons¹. *Journal of Applied Social Psychology*, 29(5), 994–1009.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables

be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373.

Roulin, N. (2015). Don't Throw the Baby Out With the Bathwater: Comparing Data Quality of Crowdsourcing, Online Panels, and Student Samples. *Industrial and Organizational Psychology*, 8(2), 190–196.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507–514.

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from Many Labs 2. *Canadian Psychology/Psychologie Canadienne*, 61(4), 289–298.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, 69, 487–510.

Slaney, K. (2017). Construct validation: View from the “trenches.” In *Validating Psychological Constructs* (pp. 237–269). Palgrave Macmillan UK.

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014). Data collection in the digital age: Innovative alternatives to student samples. *MIS Quarterly*, 38(2), 355–378.

Steinmetz, H. (2013). Analyzing Observed Composite Differences Across Groups. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences*, 9(1), 1–12.

Strickland, J. C., & Stoops, W. W. (2019). The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Experimental and Clinical Psychopharmacology*, 27(1), 1–18.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of*

primary traits (Vol. 274). University of Chicago Press.

- Weigold, A., & Weigold, I. K. (2021). Traditional and Modern Convenience Samples: An Investigation of College Student, Mechanical Turk, and Mechanical Turk College Student Samples. In *Social Science Computer Review* (p. 089443932110068). <https://doi.org/10.1177/08944393211006847>
- Winton, B. G., & Sabol, M. A. (2021). A multi-group analysis of convenience samples: free, cheap, friendly, and fancy sources. *International Journal of Social Research Methodology*, 1–16.
- Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33(1), 42–57.
- Yoon, M., & Lai, M. H. C. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(2), 201–213.
- Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: threatened morality and physical cleansing. *Science*, 313(5792), 1451–1452.
- Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *The Behavioral and Brain Sciences*, 41, e120.

Study Design Table

Question	Sampling plan	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
RQ1. To what extent do measures function equivalently across different convenience samples in the Many Labs projects?	<p>Using the previously collected open data from the Many Labs projects, we will examine every measure that meets our criteria for baseline model fit.</p> <p>We will use only data from participants collected in English.</p>	<p>We will test the equivalence of loadings (metric equivalence) and intercepts (scalar equivalence) using likelihood ratio tests for each measure and sample group pair examined at $\alpha = .05$. If the equivalence of all loadings or intercepts is rejected, we will test the equivalence of parameters at the item level using univariate score tests at $\alpha = .05 / \text{the number of items}$. We will also calculate and report dMACS effect sizes at the item level.</p>	<p>According to our review of the simulation literature on the likelihood ratio test for detecting measurement non-equivalence, we most likely have power of 80% or greater for tests involving only the 9 largest samples we are examining. Tests involving the 5 smaller samples may be underpowered and results will be discussed with caution.</p>	<p>If all measures are equivalent across all convenience samples: these samples are likely to display measurement equivalence. The pooling of samples in the ML was justified, and pooling or comparing measurements using others samples from these sources without correcting for non-equivalence is likely to be justified in future cases, though not guaranteed.</p> <p>If some measures are equivalent across convenience samples but others are not: measurement equivalence for convenience samples is dependent upon the construct and/or the specific measure. It should be tested or accounted for if measures from these data sources will be pooled or compared.</p> <p>If some crowdsourced samples are equivalent with student samples and others are not: measurement equivalence across convenience samples is dependent on the specific source,</p>	<p>The theory that measurement properties are equivalent across convenience sample sources (student and crowdsourced). This theory is assumed by the pooling of these data sources using uncorrected sum scores in the ML projects.</p>

2 CONVENIENCE SAMPLES AND MEASUREMENT EQUIVALENCE

				<p>rather than being generalizable across crowdsourced and student samples more broadly. Interpretation will depend on the pattern of results. Given the sample from India, language and culture may be a more reliable source of non-equivalence than convenience sample type.</p> <p>If all measures are non-equivalent across all convenience samples: data from these sample sources should not be pooled or compared without considering potential measurement differences, as they are likely to be a reliable source of non-equivalence. Pooling these samples was not justified in the ML and may have impacted results.</p>	
RQ2. When measures are non-equivalent, does correcting for this change the statistical significance or effect sizes of the replications?	Based upon the analyses conducted for RQ1, we will examine for RQ2 only the measures and samples which demonstrate configural equivalence but display statistically significant	We will develop a partial equivalence model for each measure and sample pair on the basis of the results of the univariate score tests from RQ1. This model will restrict parameters found to be equivalent so they are equal across groups and free parameters that display statistically significant non-equivalence. We will generate factor scores from this multiple group model,	Answering this research question will itself constitute a sensitivity analysis. We are not attempting to make inferences to other cases with these analyses; rather, we are aiming to describe whether the presence of measurement non-	<p>If the results of the replications are not changed by correcting for non-equivalence, then, while the pooling of the samples was not justified in the cases where they displayed non-equivalence, the results were robust to this.</p> <p>If the results of the replications are changed by correcting for non-equivalence, then these findings are not robust to the presence of non-equivalence. This may serve as a cautionary note and impetus for changing research practices of researchers pooling or comparing samples from these sources, although the results will not necessarily generalize to other cases, as the robustness of findings depend on particular</p>	This analysis is not attempting to disprove any theory, but rather explore the robustness of the ML findings to the presence of measurement non-equivalence.

3 CONVENIENCE SAMPLES AND MEASUREMENT EQUIVALENCE

	metric or scalar non-equivalence.	which will correct for the non-equivalent parameters. We will reproduce the replication effects using these factor scores and compare these results to the effects estimated using original scoring methods. To determine whether effect sizes are different, we will calculate 95% confidence intervals.	equivalence has had an impact on the estimation of effects in the ML replications.	features of the data in each case.	
--	-----------------------------------	---	--	------------------------------------	--