A Problem in Theory and More: Measuring the Moderating Role of Culture in Many Labs 2

Robin Schimmelpfennig[1*], Rachel Spicer[2*], Cindel J. M. White[3], Will Gervais[4], Ara Norenzayan[5], Steven Heine[5], Joseph Henrich[6], and Michael Muthukrishna[2,7]

[1] Faculty of Business and Economics, University of Lausanne

[2] Department of Psychological and Behavioural Science, London School of Economics and Political Science

[3] Department of Psychology, York University, Canada

[4] Centre for Culture and Evolution, Psychology, Brunel University London

[5] Department of Psychology, University of British Columbia

[6] Department of Human Evolutionary Biology, Harvard University

[7] Canadian Institute for Advanced Research (CIFAR), Canada

[*] Shared first authors

Corresponding authors:

Michael Muthukrishna:	m.muthukrishna@lse.ac.uk
Robin Schimmelpfennig:	robin.schimmelpfennig@unil.ch
Rachel Spicer:	r.a.spicer@lse.ac.uk

Abstract

The multi-site replication study, Many Labs 2 (ML2), attempted to test whether population, site, and setting variability moderates the likelihood of replication and effect size. The analysis concluded that sample location and setting did not substantially affect the replicability of findings. In this paper, we raise several issues with the ML2 approach to adjudicating the effect of culture that cast doubt on this conclusion. These theoretical and methodological problems (pre-registered at https://osf.io/6exr4) involve the: (1) selection of studies and sample sites for replication that are not theory-driven, (2) sampling of mostly WEIRD people around the world, (3) conflation of participants' cultural backgrounds with the country where the samples came from, (4) use of the WEIRD backronym by decomposing it into a scale, and (5) application of a mean split to this WEIRD variable. Moreover, simulations reveal strikingly low statistical power for detecting cultural influences in a multi-side study designed like ML2. We propose methodologies to address problems (3) to (5) by re-analyzing the ML2 dataset using an alternative approach. These results suggest that tackling only some of the design problems is insufficient to overcome the underlying theoretical and methodological deficiencies. We conclude with specific recommendations for assessing the role of population variability in future multi-site studies that address evidentiary value and effect size.

Keywords: WEIRD, cultural differences, Many Labs, cultural evolution, heterogeneity testing

A Problem in Theory and More: Measuring the Moderating Role of Culture in Many Labs 2

Many psychological findings do not replicate well (Camerer et al., 2018; Open Science Collaboration, 2015), likely due to both methodological malpractice and a lack of robust theory (Gervais, 2021; Munafò et al., 2017; Muthukrishna & Henrich, 2019). In the past decade, we have learned much about the reasons behind these replication failures, which have prompted methodological reform and the development of 'best practices' (Munafò et al., 2017; Nelson et al., 2018; Nosek et al., 2022).

In a separate vein, there has been growing awareness in the field that our knowledge of human behavior is heavily skewed by an empirical dataset overwhelmingly composed of people from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Apicella et al., 2020; Henrich, 2020; Henrich et al., 2010; Muthukrishna et al., 2020, 2021). But a large and growing body of contemporary cultural evolutionary theory and empirical data reveals that humans are a cultural species, evolved to be contextually and culturally embedded decision-makers. That is, their cultural backgrounds affect how they think, feel, and reason (Heine & Norenzayan, 2006; Henrich, 2020; Nisbett et al., 2001). This makes it problematic to build a behavioral science from any single population.

While both problems – replication failures and culturally homogenous sampling – have received widespread attention, surprisingly little is still known about the role of population diversity in the replicability and the effect sizes of psychological findings. For our understanding of human behavior to generalize beyond WEIRD cultural psychology published studies must be replicated with cross-cultural samples to demonstrate the robustness and limitations of psychological effects.

A high-profile research project attempted to address the moderating role of population variability (Klein et al., 2018). In the multi-site project Many Labs 2 (ML2), Klein et al. (2018) ran 28 classic and contemporary research studies distributed over 125 sample sites, comprising 15,305 participants in 36 countries. They found that 14/28 effects (50%) showed a statistically significant effect ($p < .0001$) in the same direction as the original study (15 effects replicated with the common threshold of $p < .05$). In a pre-registered design, ML2 tested whether the 28 included effects varied across different contexts (e.g., paper/pencil vs. computer-based, different cultural contexts). To investigate cultural variation as a potential explanation for heterogeneity and non-replication across samples, the researchers tested whether each effect was moderated by a binary "*WEIRDness*" scale (which was not pre-registered). We refer to it as "*ML2-WEIRDness*" in this manuscript to distinguish between scale and original backronym (Henrich et al., 2010). For each sample site, the cultural background was determined, identifying the "cultural context" of participants based on the country the sample was situated in. The *ML2-WEIRDness* score itself was calculated by decomposing the backronym into its 5 constituent letters, aggregating these scores, and taking a mean split to partition *WEIRD* from non-*WEIRD* countries.

The heterogeneity of samples, comparing Klein et al.'s classifications of *WEIRD* vs non-*WEIRD*, was calculated using the *Q*, tau, and I^2 measures (Borenstein et al., 2011). The authors found few moderating effects of the *ML2-WEIRDness* scale. After correcting for multiple comparisons, Klein et al. found evidence that in 3 of 28 replicated studies, the effects were significantly moderated by *ML2-WEIRDness* (namely: Huang et al., 2014; Knobe, 2003; Norenzayan et al., 2002). The summary of the study's results in the abstract asserts that: "*Exploratory comparisons revealed little heterogeneity between Western, educated, industrialized, rich, and democratic (WEIRD) cultures and less WEIRD cultures (i.e., cultures with relatively high and low ML2-WEIRDness scores, respectively).*"

The project represents a laudable effort. However, we identify several, potentially fatal, problems with the study design and methodology. These problems include (1) the selection of studies and sample sites for replication, (2) sampling mostly WEIRD people around the world, (3) conflating participants' cultural background with the country where the samples came from, (4) treating the WEIRD backronym as a theory by decomposing it into a *ML2-WEIRDness* scale, and (5) using a mean split of that *ML2-WEIRDness* variable.

Our article offers four contributions.

1) We state our concerns about the ML2 study design's ability to document cross-cultural variation in psychological effects; these are common problems that future cross-cultural multisite studies must address to appropriately and effectively assess whether population variability moderates the replicability and size of psychological effects.

2) We bolster the implications of ML2's sampling decisions by simulating an ML2-like environment and assessing the degree to which there is sufficient power to test cross-cultural effects.
3) We show how some of the concerns we observe in ML2 can be solved with an improved methodological approach. We illustrate this alternative methodological approach by re-analyzing ML2 data based on a pre-registered multi-step protocol (https://osf.io/6exr4) (e.g., using an inductive, empirically-driven measure of cultural difference).
4) We synthesize the implications of the methodological problems, the simulation approach to detect statistical power, and the pre-registered re-analysis of ML2, to offer a set of guidelines and recommendations for more theoretically-motivated, high-powered multi-site investigations of cultural differences in the future.

Problems in theory and methodology

We identify several problems – theoretical and methodological – with the way ML2 tested for the moderating role of population variability. We argue that flaws in ML2's sampling choices, atheoretical design, and low statistical power (when it comes to moderation by sample) reduced its ability to detect potential cultural moderation. In this section, we describe the issues, which we pre-registered ahead of the re-analyses. Next, we simulated the effective statistical power that the ML2 study possessed to detect cross-cultural moderators. Following that, we attempted to see if an inductive, empirically-driven approach based on cultural differences among populations could improve the analysis (it did not). The problems in ML2 are as follows:

Absence of any theory in the selection of studies and sample sites for replication.

Klein et al. (2018) did not explain their theoretical basis for selecting the studies to be replicated and did not provide theoretically-grounded predictions regarding which psychological effects should and shouldn't generalize cross-culturally. This is puzzling, given the rich theoretical literature on cross-cultural variability in psychology (Nisbett et al., 2001; Schulz et al., 2019; Talhelm et al., 2014). For example, they chose to include the study by Huang et al. (2014), who experimentally explored cultural differences in metaphoric associations of living in the north or the south of a city. However, the ML2 project then expanded the list of the sampled populations beyond the original populations (US and Hong Kong), without considering the explanation for why these specific cardinal directions might lead to different metaphorical associations across populations, especially because there is no reason to expect that north-south economic differences would be geographically universal (e.g. Canada is unlikely to fit this pattern).[1] Despite considering this potential problem in their pre-registration, it was not obvious from the available code how and why this north-south dimensionality was expanded to other countries in the analysis.

In other cases, the initial cultural context of a study was altered without theoretical reasoning about expected variation in the new context. For example, Norenzayan et al.'s study (2002) found cultural

[1] While Klein et al. did conduct subset analyses, including only participants from the US and Hong Kong respectively, for their main analysis they included the entire sample.

differences in a highly standardized sample of US college students who were very similar (i.e., matched based on cognitive abilities and education) except for their cultural background, while ML2 expanded the test to variation across countries, without considering the author's original reasoning in standardizing the design (Norenzayan, 2018). It is possible that the effects were selected based on theoretical foundations, but we could not find these stated in the article or the pre-registration.

Similarly, the selection of sampling sites does not seem to have been guided by hypotheses about meaningful cultural variation. It is regrettable and ironic that a Many Labs study, designed to assess the role of sampling heterogeneity, appears to have made sampling choices at random, or more likely, by convenience. The strongest possible test of cultural variation of a particular phenomenon would require sampling from populations that are known to vary maximally on a theoretically relevant dimension (Norenzayan & Heine, 2005). Without a sound theory to explain the source of cross-cultural variation, it is difficult to know the range of cross-cultural psychological differences represented by these sites, and this necessarily weakens any conclusions that can be drawn about any particular effect's cultural variability. The selected effects or sample sites cannot be accounted for retrospectively. Future tests of cultural moderation that are grounded in theory would be far more convincing and a stronger contribution to the literature.

Crucially, using conveniently available samples is not a harmless choice because, given the cultural background of the lead authors and the structure of their social networks, this can produce wildly biased sets of populations—which appears to be the case here. Of course, psychologists are familiar with the representative heuristic, which in this situation will lead many readers to implicitly assume that this convenience sample is roughly equivalent to a representative or random sample. It is not, as we will demonstrate. As a rule, authors should explicitly justify and defend how they selected their populations, and state which factors apart from culture are likely to vary between those populations.

The sample consisted mostly of WEIRD people around the world

Setting aside the lack of theoretical considerations in the sampling methodology of ML2, what is even more worrisome is that the ML2 subject pool consisted of participants who are predominantly US-based and overwhelmingly WEIRD (see Figure *1* for sample composition of ML2's Slate 1 and Slate 2). Indeed, only a fraction of the participants was obtained from non-Western populations. 39% of all participants were sampled from the US – the WEIRDest of all countries.

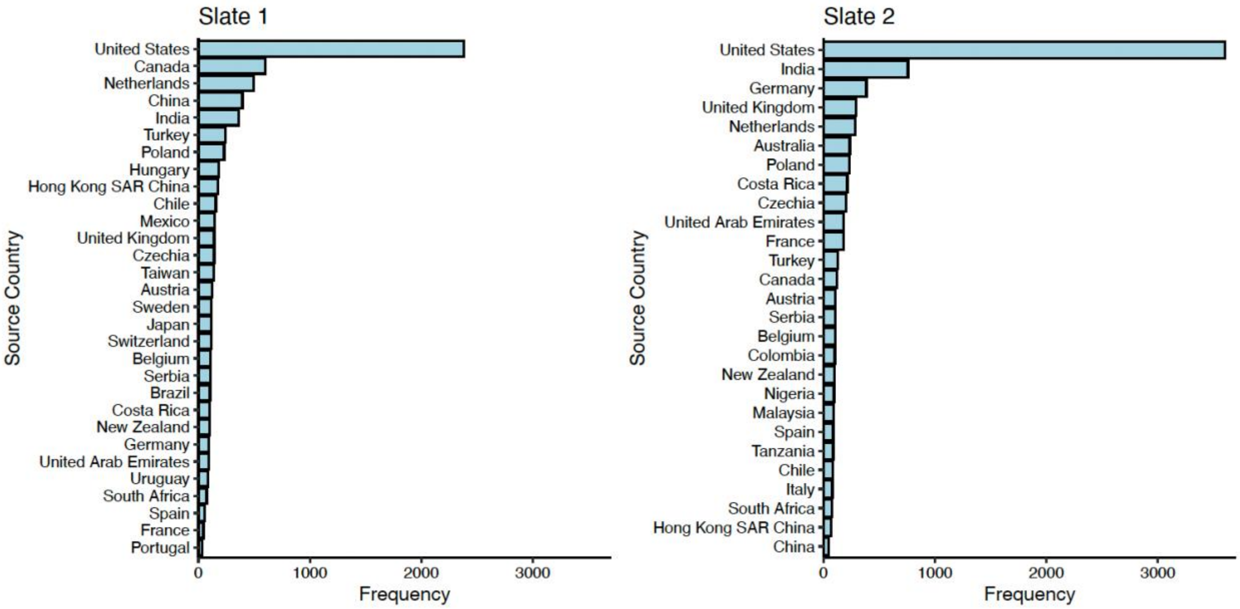

Figure 1. Frequency of participants per source country in which the participants were sampled. The red dashed line indicates the minimum number of participants included in any one sample: 36.

We can be more precise about the extent of this problem. Cultural fixation (CF_{ST}) provides a continuous measure of cultural similarity between groups across a range of cultural traits (Muthukrishna et al., 2020). Because the United States is a country at the extreme end of the WEIRD spectrum, a country's cultural distance from the US can be used as an index of how WEIRD that sample is. In the ML2 data, the average CF_{ST} distance from the United States using participants as the unit of observation is 0.062; using sample sites as the unit of observation like in ML2, it is 0.055. For comparison, this is smaller than the cultural distance between the United States and Germany (0.069). Overall, there is scant cultural variability in the ML2 sample.

Furthermore, despite cross-*country* variation in sample sites, the sites were based at universities and Amazon MTurk, which is known to oversample from high SES populations (Berinsky et al., 2012). Thus, the samples are likely to be skewed towards participants who are high SES, highly educated, and digitally literate, and therefore much more likely to be Westernized and at the "WEIRD" end of the distribution within each society. Just as sharing a religious affiliation predicts cultural similarity among people living in different countries (White et al., 2021), sharing a high socioeconomic status, and participating in WEIRD institutions, such as higher education institutions may likewise drive cultural similarities between people across nations. Of course, we recognize the practical and financial constraints that researchers face when it comes to recruiting globally diverse samples; nevertheless, the strength of inferences about generalizability across populations is proportional to the extent and scope of diversity that is captured in sampling choices. The ML2 dataset largely consists of highly educated individuals from Western populations, rendering it poorly suited for investigating the cultural variability of psychological effects and unsuited for making big claims.

Conflating cultural background with sample country

The issue of sampling WEIRD people is further exacerbated by calculating the *ML2-WEIRDness* score for each sample site, based on the country of origin of that site, irrespective of the original country of origin of participants. Clustering by the source country of the sample site rather than by the individuals' origins conceals potential psychological variation including the possible migration background of participants. Indeed, the samples had significant shares of migrants (e.g., international students) at some sites (up to 61% in the UAE and 45% in Canada; see Tables S7: 8). Figure *2* shows the constitution of birth countries for participants in the different source countries. Strikingly, many participants from the US were born in other countries, indicating the possibility of cultural variation hidden by the approach taken by ML2.

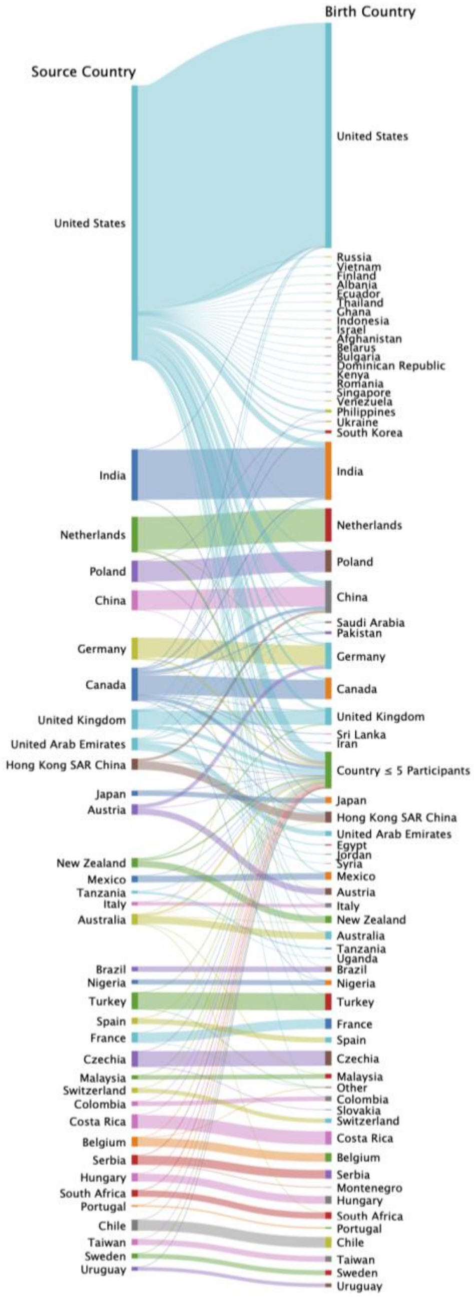

Figure 2. Constitution of sample countries by birth countries as indicated by participants. Source countries are shown on the left and birth countries are on the right (A detailed overview including birth countries with less than 5 participants can be found in Table S6).

Assessing cultural moderation by decomposing the backronym (WEIRD), a rhetorical device

We were surprised to see that Klein et al. assessed cultural moderation by decomposing the letters of the WEIRD backronym. We are not aware of any conceptual, empirical, or theoretical justification for this move (and the authors of ML2 did not provide one). Indeed this is inconsistent with the formulation of the WEIRD people problem (Henrich et al., 2010). The catchy backronym captures some aspects of the regions and demographics that are overrepresented in the psychological record – namely, being Western, Educated, Industrialized, American, and high SES (often students) – but it hardly captures the defining characteristics or mechanisms driving differences between these societies. WEIRD was designed as a consciousness-raising device aimed at reminding experimental behavioral scientists about psychological diversity (Apicella et al., 2020), not as a theoretical operationalization of the explanatory concept. In their introduction, Henrich et. al. explain:

> We emphasize that our presentation of telescoping contrasts is only a rhetorical approach guided by the nature of the available data. It should not be taken as capturing any unidimensional continuum or suggesting any single theoretical explanation for the variation. (2010, p. 62)

The point is that Henrich et. al. were explicit that the WEIRD backronym is not intended to embody or summarize any key theoretical or conceptual factors important for explaining global psychological variation. This would turn a mnemonic conscious-raising device into a theoretical construction. Of course, one could take a purely inductive, empirically-driven approach based on cultural differences among populations–without bestowing any theoretical import on the backronym's letters–to generate such a measure. This is the approach we take, drawing on earlier work by Muthukrishna et. al. (2020).

Mean Split of WEIRD scale

A further issue is that the authors ascribed binary codes (1 and 0) to sample sites based on a mean split and thus sum up all WEIRD, and especially non-WEIRD samples under the same category. It is questionable whether this dichotomy is informative. A more precise, continuous coding of cultural distance between samples would be more appropriate. These shortcomings are obvious when one looks at the ML2 coding of their samples, which results in rather surprising binary *ML2-WEIRDness* values. For example, the sample site at the American University of Sharjah in the UAE with its gold-plated pillars is coded as non-rich; Chile was coded as categorically the same as Germany and Sweden, but categorically different from near-neighbors Spanish-speaking Costa Rica and Uruguay. South Africa was coded as categorically like China and India, but categorically different from other Commonwealth states such as Australia and New Zealand. This lack of face validity is also a sign of the adverse effects of missing theory about culture and population variability (Gervais, 2021). The acronym-based operationalization of cultural difference reveals a highly skewed distribution, in which samples were overall quite *ML-WEIRD* and the *ML-WEIRD* sample sites were culturally very similar. In contrast, the non *"ML2-WEIRD*" samples had larger variations. Figure 3a shows the distribution for *ML2-WEIRDness,* and the applied mean split, on the country level.

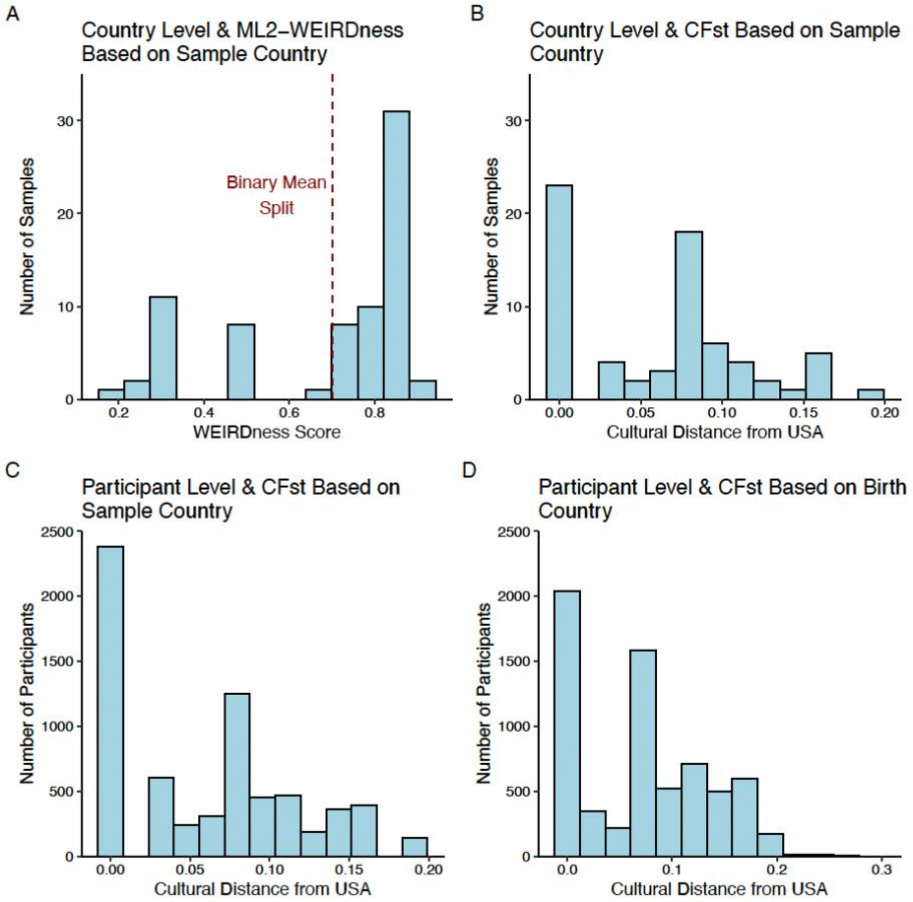

Figure 3. Frequency distribution for different measures of cultural differences. Figure 3a shows the distribution of ML2-WEIRDness score in ML2 (data retrieved from Klein et al. 2018 with cutoff for the binary mean split at 0.7) on the sample level, based on the source country. 3b shows the distribution of CF_{ST} values (Muthukrishna et al., 2020) on the sample level, based on the source country. 3c shows CF_{ST} distribution on the participant level, based on the source country, and 3d shows the CF_{ST} values on the participant level, based on the birth country of participants. High values on the ML2-WEIRDness score indicate comparably "WEIRD" countries, while high values for CF_{ST} indicate comparably "non-WEIRD" countries, The histograms show the data for Slate 1. A complete set of histograms including data from Slate 2 can be found in Figure S3.

Together, Problems 3), 4), and 5), that is, conflating cultural background with sample country, an atheoretical operationalization of the *ML2-WEIRDness* scale, and the mean split motivated us to rerun the analysis with a different measure of cultural differences and identifying culture also by the birth country of participants.

Measuring Cultural Distance with Cultural FST

As mentioned above, we apply an inductive measure of cultural distance that could be used to assess the effect of sampling variability. Muthukrishna et al. (2020) developed a tool to measure the degree of similarity between the cultural values, beliefs, and practices of different groups of people. This measure provides an empirically-driven way to quantify worldwide variability in culture, by measuring the overall cultural distance between any two countries for which data is available (Muthukrishna et al. used the 2005 and 2010 waves of the World Values Survey (Inglehart et al., 2014; Muthukrishna et al., 2020)). CF_{ST} as implemented in our re-analysis focuses on the distance between each country and the United States, the WEIRD country par excellence that is vastly overrepresented in the psychological literature (Apicella et al., 2020; Muthukrishna et al., 2020; Thalmayer et al., 2020), as well as in ML2. Figure 3b-d show the cultural distance score operationalized by CF_{ST} (Muthukrishna et al., 2020) on the country level based on where the participants were sampled (3b), and the participant level with cultural background identified by the country of the sample (3c) or the birth country of participants (3d)).

Figure 4 below shows the variation in distance from the United States in the ML2 data and its correlation to the *ML2-WEIRDness* scale. The samples come from a highly restricted range of countries, with 91% of the samples having a score of $CF_{ST} < 0.15$, even though worldwide cultural distances from the United States extend to approximately 0.3 using data from the 2005 and 2010 World Values Survey data (Muthukrishna et al., 2020).

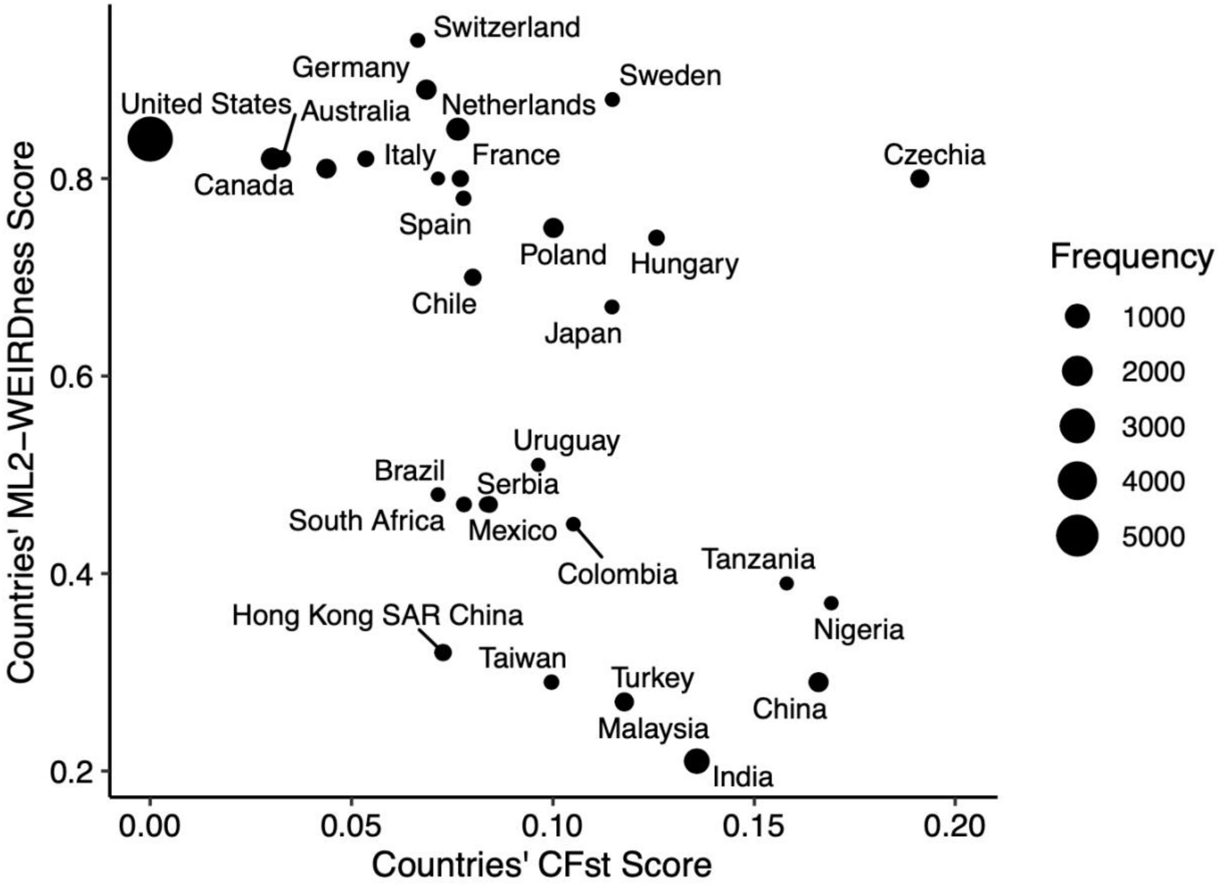

Figure 4. Scatter plot for CFST values and ML2-WEIRDness scores. The plot reveals a moderate correlation between the CFST values and ML2-WEIRDness scores before the mean split, r = 0.49.

The cultural distance scale is a more defensible proxy for a WEIRD scale (Muthukrishna et al., 2020) because it is empirically derived, but attempting to re-run the analyses with this improved scale puts us in a bind. The extreme skewness of the cultural location of the samples (see Figure *1*) means that the data is biased toward comparably WEIRD countries. To better understand what level of cultural heterogeneity and statistical power would need to be present for an analysis as the ML2 to detect an effect, we ran a simulation.

SIMULATING THE MODERATING ROLE OF CULTURE

To calibrate expectations for what conclusions one could reasonably expect to emerge from ML2's design choices, and thus also for a possible re-analysis of their results, we conducted a series of simulation studies. The overarching goal was to closely mirror the design, sampling, and analyses of ML2 (in terms of selected effects, samples included, and analyses conducted) while manipulating cultural influence: the degree to which culture moderates effect sizes. In making cultural influence an exogenously determined variable, we can obtain an answer regarding what degrees of cultural heterogeneity a setup such as ML2 is well-designed to detect. While the simulations mimic the characteristics of ML2, they could equally be adapted to simulate power for other multi-lab studies (see Quintana, 2023 for power calculation for meta-analyses.).

ML2's moderation analysis "*[...]revealed little heterogeneity between Western, educated, industrialized, rich, and democratic (WEIRD) cultures and less WEIRD cultures [...]*" (Klein et al., 2018, p. 446). This result can be interpreted in at least two ways. One possibility is that ML2 was well-designed, the analyses provide a strong test of cultural heterogeneity, and the results showing a small influence of culture are probably due to low rates of cultural heterogeneity. Another possibility is that culture has a large influence, and results indicate that the design and implementation of ML2 are poorly suited for detecting such differences. In essence, our simulation is trying to understand which of the two scenarios (little heterogeneity, but well-measured; unknown-to-high heterogeneity, but poorly measured) is more likely to be true, given a design like ML2.

In discussing limitations in their project, the ML2 authors do not explicitly acknowledge the possibility that their project was not designed in a manner that allowed powerful and conclusive tests of heterogeneity. One may infer that they, therefore, favor the former interpretation of ML2's results as reflecting genuinely low rates of heterogeneity resulting from a study that was well-designed to support such a conclusion. We also note that when the authors discuss statistical power, they do so in the context of detecting the main effects within each study; little to no mention is made of the effective statistical power of the crucial heterogeneity tests that are the inferential backbone of their paper. In short, ML2 acknowledges that other as-yet-untested effects might show heterogeneity, but they do not discuss the possibility that their methods and analyses might be unable to deliver evidence of heterogeneity even when it's present. Our simulations may help provide context for interpreting that key (largely null) heterogeneity tests and inform intuitions about whether ML2 reflects little heterogeneity that's been well-measured or an unknown degree of heterogeneity in a design that may be severely underpowered to detect it.

Simulation Setup

To begin with, we created a simulation environment in which multiple studies in different settings or countries are run on a given effect. Countries in the simulation were randomly drawn, in proportion to their representation in ML2 (see Figure *1*; e.g., USA samples were vastly more likely to be included than samples in, say, Uruguay or the United Arab Emirates). We created the *ML2-WEIRDness* scores, performed their same mean split into *ML2-WEIRDer* and less *ML-WEIRD* countries, and meta-analytically quantified heterogeneity exactly as they did, using *Q*, tau, and I^2 indices (Borenstein et al., 2011).

Simulations always include assumptions, and we strove to model all assumptions both transparently and quite generously. Representation of countries and effect sizes directly mapped onto ML2's design. To manipulate the levels of cultural influence in the study, we model both cultural differences (captured by CF_{ST} (Muthukrishna et al., 2020)) between countries and the moderating role of these cultural differences of the effects (captured by the variable *cultural influence*). We set the USA as the reference country for effect sizes, as it is by far the most overrepresented in ML2, and modeled culture by having the true effect size within each country diverge from the USA effect size by an amount that relates to CF_{ST} and the given cultural influence. As illustrated in Equation 1, each country's (i) effect size (d_i) would thus include the effect size of the chosen reference country USA (d_{USA}), adjusted by the country's cultural distance from the USA ($\mathrm{CF}_{\mathrm{ST\_i}}$) multiplied by a constant that reflected the degree of cultural influence (Cultural Influence) in a given simulation:

$$d_i = d_{USA} * \left(1 - \left(CF_{ST\_i} * \frac{\text{Cultural Influence}}{CF_{ST\_max}}\right)\right) \qquad \text{(Equation 1)}$$

For ease of presentation, we show simulation results for three levels of *Cultural Influence: no influence, moderate influence,* and *strong influence* (see Figure 5).

No influence ($\text{Cultural Influence} = 0$) refers to a scenario in which all countries have the same effect size regardless of their cultural differences. Simulations with no influence thus had all countries drawn from the same reference effect size (d_{USA}), ignoring any presence of cultural differences (CF_{ST_i}). That is, in the *no influence* condition, each country's CF_{ST} is multiplied by zero, entirely leveling our slate of countries to the same effective effect size for all effects (e.g., a $d_{USA} = 0.5$ effect in the USA would also be a $d_{CN} = 0.5$ in China – see Figure 5).

Moderate influence ($\text{Cultural Influence} = 0.5$) refers to a scenario where the most culturally distant countries have effects half as large as in the USA. Simulations at moderate influence thus represented the most culturally distant countries from the USA as having effect sizes *half as large* as those observed in the USA (e.g., $d_{USA} = 0.5$ in the USA would be $d_{CN} = 0.25$ in China).

Strong influence ($\text{Cultural Influence} = 1$) refers to a scenario where effects entirely attenuate in the most culturally distant countries. Simulations at strong influence thus represented effects fully disappearing in countries the most culturally distant from the USA (e.g., a $d_{USA} = 0.5$ in the USA would be a $d_{CN} = 0$ in China).

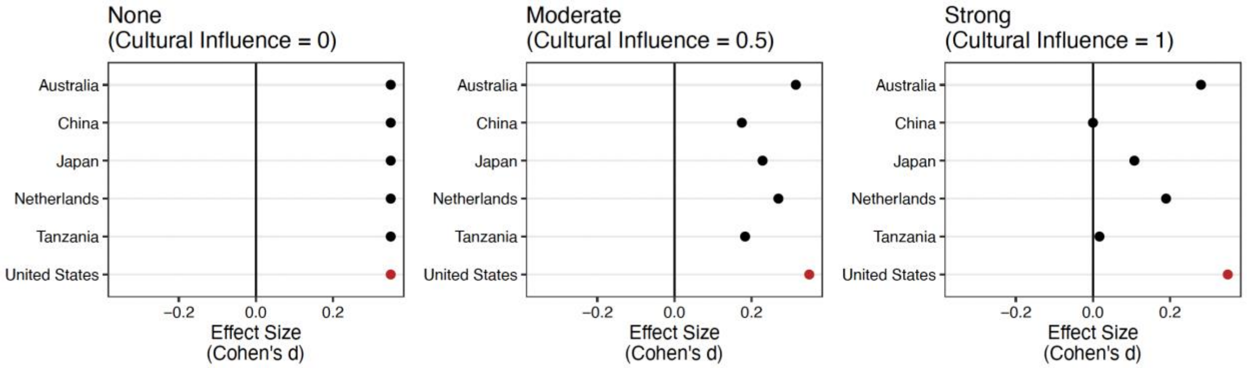

Figure 5. The effect of culture on effect sizes at different levels of cultural influence. Cultural Influence=0 means no cultural influence, Cultural Influence=0.5 means moderate cultural influence, and Cultural Influence=1 means strong cultural influence.

Simulation Results Summary

We sought to assess how much power ML2 had to detect different levels of cultural influence given a typical effect size for social psychology (for other variations of the simulation, see Supplementary). We simulated 3000 multi-site style investigations of a typical social psychology effect, each investigating a "true" d = .35 effect size, with k = 60 samples drawn from countries in proportion to their actual representation in ML2.

Our 3000 simulations consisted of 1000 simulated multi-site studies of a d = .35 effect at each of the three levels of cultural influence, as described above. For each of these simulated multi-site studies, we could perform the same analyses reported in ML2. By simulating Many Labs studies, we could directly assess the statistical power at which cultural influence could be detected, given known effect sizes and rates of cultural influence that we directly controlled.

Overall results show that statistical power is dismally low to detect cultural influence. For example, assuming a typical social psychology effect size (*d* = .35) for the simulation, power is just 33% to detect cultural influence via the moderation test of *ML2-WEIRDness*, even if the modeled cultural influence is strong (see Table 1). Put differently, the 3/28 studies (11%) found in ML2 to be moderated by *ML2-WEIRDness* would best map to a scenario in which cultural influence across all studies is at least moderate. Using Many Labs 2's criteria as benchmarks, power to detect cultural moderation was quite low, see Table 1:

Table 1. The power of the statistical measures used in ML2 to detect cultural influence for different analytical approaches. The last column titled "ML2-WEIRDness moderation" shows the power of the moderation analysis, the analysis we are replicating in the later part of the paper. Power ranges from 0 to 87%, depending on the simulated effect size. In the ML2 study, the authors found 3/28 studies with significant moderation via ML2-WEIRDness.

Level of cultural influence	***Q* test**	**tau**	***I^2***	***ML2-WEIRDness* Moderation**
None (0)	0.00	0.04	0.01	0.01
Moderate (.5)	0.01	0.1	0.04	0.04
Strong (1)	0.06	0.44	0.32	0.28
ML2 Observed Values	*0.39*	*0.32*	*0.46*	*0.11*

This simulation suggests that, for effects one would likely encounter in social psychology, ML2 was probably underpowered to detect all but the strongest levels of cultural influence. Strikingly, power to detect moderation in which effects entirely disappears in countries dissimilar to the USA *(strong influence)* ranged from 6% to a mere 41%, depending on which of ML2's chosen criteria one focuses on. For the heterogeneity analysis using the *Q*, tau, and I^2 measures (Borenstein et al., 2011), results like those in ML2 are entirely consistent with strong (or extreme) levels of actual cultural influence.

Our initial simulation results focused on the power to detect different levels of cultural influence, given a single typical effect size. But analytic performance varies across both effect sizes and degrees of cultural influence. After all, ML2 included effect sizes ranging from practically nonexistent to quite large. Our simulations show that for all criteria for detecting heterogeneity (*Q*, tau, I^2, and *ML2-WEIRDness* moderation), power to detect moderate cultural influence was poor for all but quite large effect sizes. This combination of factors means that an ML2-style investigation would have quite low power for detecting combinations of effect sizes and cultural influence that are quite plausible in the world (i.e., small-to-medium effects, with effect sizes largely attenuating in dissimilar populations). Because power to detect heterogeneity varied as a function of both the initial effect size and the degree of cultural influence – both quantities that one might hope to assess in a ML2-style project – it makes results from an ML2 investigation difficult to interpret: is a given null result on a measure of heterogeneity reflective of an actual lack of heterogeneity, or merely low power to detect whatever amount of heterogeneity is present? Our simulation results indicate that the second scenario is plausible – a problem that our re-analysis also faces.

Despite all limitations that apply to simulations to model real-world settings, these simulations give us some pause in evaluating ML2's conclusions. Across thousands of simulations in which culture mattered in degrees we could precisely control, analyses like those in ML2 usually failed to detect heterogeneity. Like a 2-condition between-subjects experiment testing a typical social psychology effect size (d = .35) with a total of only a dozen participants, Many Labs 2 might have had less than 8% power to detect heterogeneity for many of its effects[2]. Put differently, any reanalysis of the data, including our own, will be similarly constrained in statistical power. Our simulations suggest that the methods used in ML2 are severely underpowered and thus preclude solid inference about the genuine degree of heterogeneity present.

ANALYSIS AND RESULTS

With these observations and caveats out in front, in this section, we develop an alternative approach to analyze the ML2 data that addresses our stated Problems (3) *Conflating cultural background with sample country*, (4) *the ML2-WEIRDness variable*, and (5) *the Mean Split*. Unfortunately, neither an inductive measure for cultural differences nor a more accurate approach to identifying cultural background on the individual level can solve Problems (1) Lack of theory in the selection of studies and sample sites for replication, and (2) Sampling WEIRD people around the world. Moreover, as ML2 found that

[2] Power to detect d = .35 with 6 participants per condition = .08.
Power to detect moderate cultural heterogeneity (effects half attenuate) for d = .35, by *tau* criterion = .08.

many of the results of the individual studies do not have evidentiary value at all, then we should not expect these null findings to vary cross-culturally. So, at best, an improved reanalysis will extract the variation in this available data, while recognizing that these samples may lack sufficient variation and statistical power, the protocol was implemented in a way that may have introduced a lot of noise, and the chosen studies only represent a subset of psychological science. The results of our re-analysis, as stated in our pre-registration, should thus not be perceived as a final verdict on the role of cultural variation in the given effects. Answering such a question about cultural heterogeneity would require data resulting from a project specifically designed for the task. To inform future studies in this realm, we chose to illustrate how some of the problems we state can be operationalized in an improved and theory-driven methodological approach.

We analyzed the existing ML2 data per the pre-registered analysis plan as described below. Specifically, our re-analysis incorporates three changes, compared to the ML2 approach:

1. We replaced the dichotomous *ML2-WEIRDness* score with a continuous proxy for a WEIRD scale—cultural distance from the United States (Muthukrishna et al., 2020).
2. We operationalized cultural distance from the US not only at the sample site level but also at an individual level by identifying the birth country of the participants.
3. In addition to calculating the cultural distance from the US, we calculated the cultural distance between participants based on their birth country and used these between-level distances to estimate the moderating role of culture.

Figure *6* summarizes the analysis plan:

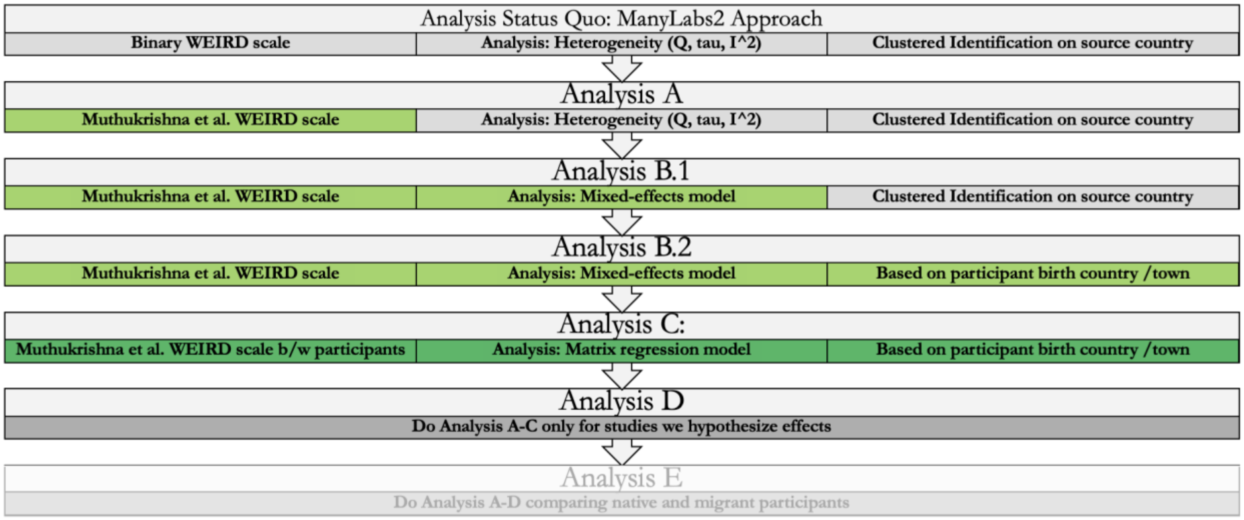
Analysis Status Quo: ManyLabs2 Approach
Binary WEIRD scale
Analysis: Heterogeneity (Q, tau, I^2)
Clustered Identification on source country
Analysis A
Muthukrishna et al. WEIRD scale
Analysis: Heterogeneity (Q, tau, I^2)
Clustered Identification on source country
Analysis B.1
Muthukrishna et al. WEIRD scale
Analysis: Mixed-effects model
Clustered Identification on source country
Analysis B.2
Muthukrishna et al. WEIRD scale
Analysis: Mixed-effects model
Based on participant birth country /town
Analysis C:
Muthukrishna et al. WEIRD scale b/w participants
Analysis: Matrix regression model
Based on participant birth country /town
Analysis D
Do Analysis A-C only for studies we hypothesize effects
Analysis E
Do Analysis A-D comparing native and migrant participants

Figure 6. Overview of the pre-registered analysis plan. The colors indicate the stepwise change in the conceptualization and operationalization of measuring cultural differences. First, we used the analytic approaches of Klein et al. (gray). We then changed to cultural distance based on Muthukrishna et al. (2020) (light green), and finally used approaches based on Muthukrishna et al. (2020), but between participants (dark green). We were not able to perform Analysis E as stated so this is grayed out.

Below, we state our research questions, the associated analysis, and the summarized results of the analyses. The inclusion criterium in the analysis was a minimum of 36 participants per sample, as this was the minimum number of participants of a sample included in the original ML2 analysis (the *uniporto* sample from Portugal), although we acknowledge that such small samples will necessarily be unreliable. We follow this haphazard threshold for convenience because the stricter criteria are even less tenable to illustrate the methodological approach because they would result in more exclusion of countries and raise issues with statistical power (see Supplementary Section 3 for results based on other inclusion criteria).

Does cultural distance measured by the Muthukrishna et al.'s cultural distance scale at a sample level explain variation in the outcomes of the studies?

Analysis: To ensure comparability, here we used an almost identical analytical approach as the authors of the ML2, except that we replaced the dichotomous *ML2-WEIRDness* score with the continuous CF_{ST}-based cultural distance score (Muthukrishna et al., 2020). As such, we ran a Random-Effects model with cultural distance as a moderator, and similarly established heterogeneity of samples, using the Q, tau, and I^2 measures (Borenstein et al., 2011). CF_{ST} values are not yet available for all countries[3], which led to the exclusion of some of the source countries in the analysis (see the Supplementary for additional analyses imputing missing CF_{ST} values).

Summary of results: We hypothesized that the continuous CF_{ST} cultural distance would be a stronger predictor of the effect size, compared to the *ML2-WEIRDness* score. Figure 7 summarizes the results in a Forest plot and shows that using CF_{ST} instead of the binary *ML2-WEIRDness* variable only marginally increases the number of significant effects. Overall, CF_{ST} consistently increases effect sizes for most effects (especially those effects that replicated in ML2; bolded in Figure 7), but likely failed to reach the chosen significance level because of constraints in statistical power. As per the simulation results above, the resulting pattern can be interpreted as preliminary evidence that moderate

[3] In total 5/36 source countries were missing CF_{ST}: Austria, Belgium, Costa Rica, Portugal, and the United Arab Emirates.

levels of cultural influence likely exist, but just replacing the *ML2-WEIRDness* variable with CF_{ST} does not overcome the underlying issues in statistical power.

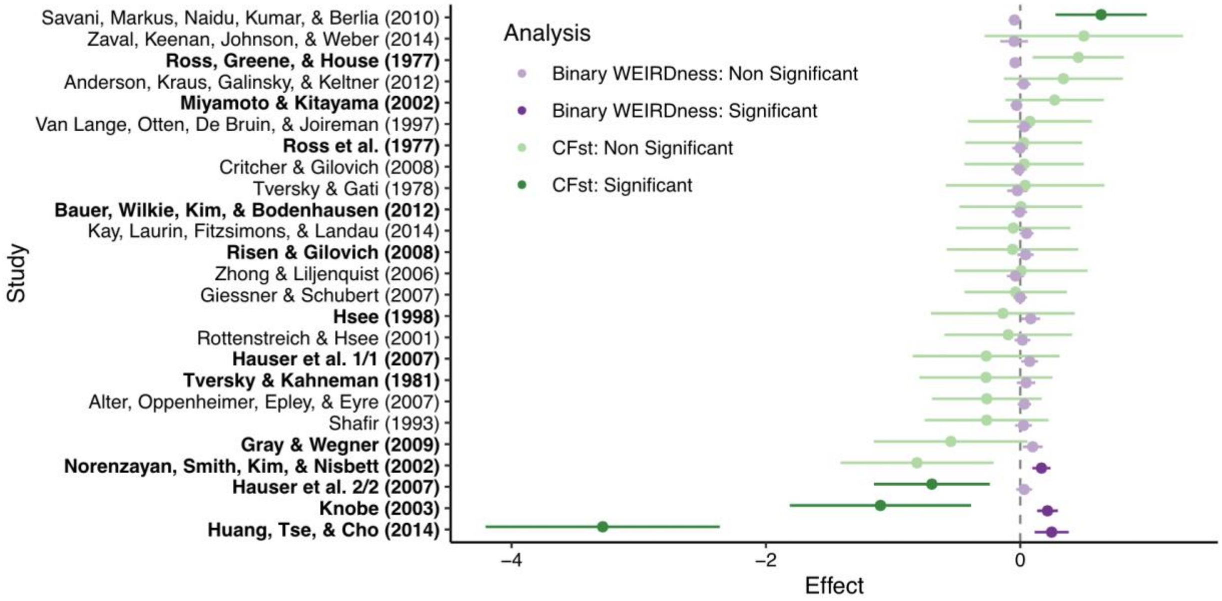

Figure 7. Summary results for Analysis A. Forest plot showing the moderating role of CFST (green) and the ML2-WEIRDness variable (purple) on effect sizes. This shows that CFST has a significant moderator role in four studies (dark green) and that ML2-WEIRDness had a significant moderator role in three studies (dark purple). The significance threshold for the migration status analysis uses the alpha = 0.004 (Slate 1) and alpha = 0.003 (Slate 2) as calculated in the original ML2 analysis using Bonferroni correction for multiple comparisons. Studies in bold replicated in the original ML2 study.

Does cultural distance identified on the individual level by the Muthukrishna et al. WEIRD scale explain variation in the behavior of participants?

Analysis: Here we used a multilevel model (MLM) to predict effect sizes. For this analysis, we relied on the variance of cultural distances to the US at the sample site level (**Analysis B1**) and in a second approach also on the individual level, identifying CF_{ST} by the birth country of participants (**Analysis B2**).

Summary of results: Analysis B1 shows a significant moderation of the *CF_{ST}* by sample site for two effects, with only one of them having been successfully replicated in ML2. Analysis B2 shows no study being significantly moderated by *CF_{ST}* by birth country. Thus, the MLM regressions in Analysis B find less significant effects of *CF_{ST}* than in Analysis A and the original analysis in ML2. There are several possible explanations for this result, including the type of models used or the relationship between cultural distance and the effect sizes not being linear. Due to our inclusion criteria of having at least 36 observations (the minimum sample size at ML2), many birth country sites were excluded from the analysis which exacerbated the power issues. In total 1735 participants from 153 countries were excluded from analysis B2, removing much cultural variation (36 countries were retained in the analysis). We thus ended up with more sample locations than birth countries, which prevented us from running a pre-registered robustness check in which the sample site of the individuals was included as a random effect (for details on the results, see Supplement Section 3.2.2.

Does the cultural distance between participants at an individual level explain variation in the behavior of participants?

Analysis: For this analysis, rather than the distance from the US, we used the direct difference in *CF_{ST}* between countries. We used matrix regression models with the same samples as Analysis B1 and B2 to assess whether individual-level differences between participants explain variations in their behavior.

Summary of results: Overall, we did not find any study with a consistently significant effect for *CF_{ST}* in the matrix model. Similar to analysis B, this is likely explained by constraints in the power, as both analyses use the same samples. A detailed summary of the analysis and its results can be found in the Supplementary.

For each study, we preregistered whether we expect culture to matter or not. We ran our analyses on this subset of culturally relevant studies as well as the full set of studies (see Supplementary Section 2.6)

Analysis: We conducted Analyses A-C only on the subset identified in Table S3. Here, we first included studies we expect to cross-culturally vary. Next, we also included studies that may vary with some caveats and go/no-go exclusion criteria described in the comments. We also ran an exploratory analysis on all studies that were replicated in ML2. For each study, we have stated whether we expect CF_{ST} to matter or not ex-ante (see Supplementary for details).

Summary of results: The summary of the analysis (for details see Table S9) suggests that given the existing dataset, Analyses B and C were not suitable to detect the moderating role of CF_{ST} in the given setting. While we stand by the theoretical critique in the pre-registered analysis, such an approach would require a larger sample size per country. For example, Analysis B2 identifies the cultural difference on the birth country level, and given the diversity in birth countries and the minimum sample size of 36 per country, much of the data was excluded (i.e., in total 1735 participants from 153 countries were excluded, 36 countries were included). Analysis C is similarly underpowered, and thus found little evidence for cultural difference.

How similar is the behavior of participants in their native country to participants not in their native country (i.e., how different are migrant populations in their country of origin)?

Analysis: We pre-registered to compare the behavior of the native and migrant participants for Analysis A-D. However, we were unable to run these analyses, as only the largest (and often WEIRDest) countries reached minimum sample sizes for inclusion (e.g., as per the above inclusion criteria of 36 participants – see detailed overview in Supplement Section 2.5.5). Therefore, to better understand whether migration status has some impact on behavior, we performed an explorative analysis of whether there were differences in behavior between migrant and non-migrant participants ignoring the potential effects of CF_{ST}. That is, we re-ran the first-stage ML2 regression models and simply added 'migration status' as a control variable. Participants were marked as having a migration status when their stated birth country differed from the source country site where data was collected. We decided to focus on samples situated in the US for our explorative analysis, as it had the largest sample size among those countries, and also serves as a reference country for the CF_{ST}.

Summary of results: Overall we found that the migration status of participants had a significant moderation effect on the behavior in 12/23 of the studies ML2 ran in the US (see Figure 8). The results of this exploratory, not pre-registered analysis cherry-picking one country should not be considered as an overall test of whether migration status matters but shows that classifying cultural background simply based on the source country where data is collected may neglect some rich cultural variation.

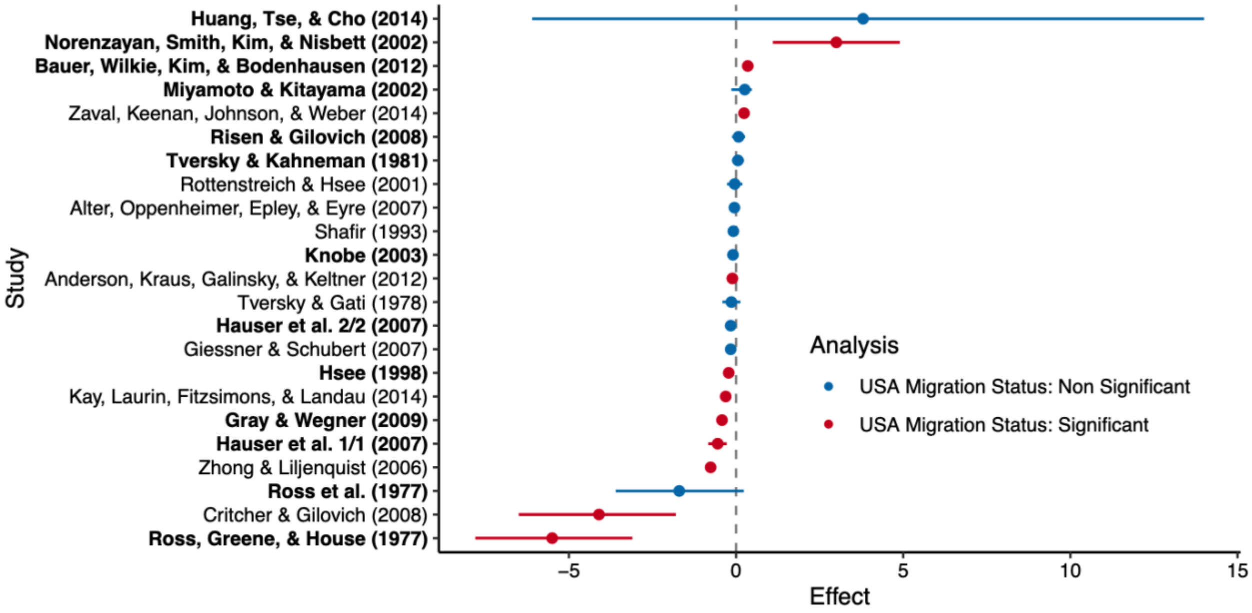

Figure 8. Summary results of Analysis E only for sample sites in the USA. Forest plot showing the effect size of migration status, which we added as a control variable to the original ML2 regression models. The significance threshold for the migration status analysis uses the alpha = 0.004 (Slate 1) and alpha = 0.003 (Slate 2) as calculated in the original ML2 analysis using Bonferroni correction for multiple comparisons. Studies in bold replicated in the original ML2 study.

Discussion

Our pre-registered approach to improving the analytical protocol for detecting the impact of cultural distance from the US showed some limited evidence that there may indeed be more cultural variance than detected in ML2. Operationalizing cultural distance via CF_{ST} improved the analysis and based on the results from the simulation, the findings would be most consistent with at least moderate levels of cultural influence in the ML2 sample. While using CF_{ST}as a measure of cultural distance fails to reach conventional significance levels for many studies, potentially because of the constraints in power, it shows increased effect size across most of the studies compared to *ML2-WEIRDness*.

Furthermore, Analysis E shows convincing evidence that the cultural variation in a given sample site is much larger if participants within a (sample) country are not lumped into the same cultural bracket, but, for example, their migration history is respected when measuring cultural background.

However, this reanalysis cannot provide a satisfactory solution to measuring the moderating role of culture on replication success. That is, the re-analysis merely addresses one small aspect of the project – an issue with the analysis. For the issues stated above the collected samples provide a comparably weak test of cross-cultural universality in effects. Any analyses using these data are therefore marred by the same serious design and sampling issues, which means that they cannot be the best test of the degree to which cultural differences matter for these seemingly arbitrarily selected psychological findings. And therefore, any attempts to improve any aspect of the analyses are limited by the range of the data and may therefore do little to change the results. Statistical analysis is a tool, and the raw materials one feeds a statistical model are at least as important as the modeling choices employed.

A more principled approach to testing the moderating role of cultural differences includes careful planning in both the design and analysis phases. In the simulation and the re-analysis, we addressed some of the existing issues in the analysis stage and suggested practical ways to alleviate them. We hope that these above chapters prove to be practical as a guideline for future studies in this realm. Nevertheless, improvements both at the design and the analysis stage are needed. Based on our above reasoning, we suggest several points that need to be considered in the different stages of such a project.

Culture may not moderate all aspects of psychology

The selection of studies for a multi-site project should be informed by the underlying theories tested (Stroebe, 2019). Similarly, we suggest that effects for replication in a study testing the influence of sampling diversity should be selected based on theoretical predictions about whether we would expect the effects to be moderated by culture. For example, for some effects, we would not expect successful replication across populations (e.g. North/South differences in socioeconomic status would not be expected to replicate across countries with different geographic patterns of wealth like Canada). Other effects may have clear theoretical or empirical evidence that suggests less heterogeneity across populations (e.g. despite differences in social norms between countries, children seem to have a uniform tendency to respond to some novel social norms across societies (House et al., 2020)). Not all aspects of human psychology/behavior are equally likely to vary across populations (Henrich et al., 2010). To understand which effects may, and may not be moderated by culture, we need to invest more effort into developing better theories for human psychology and behavior.

Selecting sample sites based on predictions of meaningful cultural differences

Sample sites should be selected based on a prediction about whether we would expect the cultural differences to produce significant psychological differences. We appreciate that coordination and availability constrain multi-site projects in their ability to selectively target sample-specific sample sites. But more statistical power in this setting does not necessarily require sampling more people from where they are available, but could also mean deliberately sampling at different sample sites chosen to reflect theoretical expectations of relevant cultural differences. That is, a more limited set of better, more theory-driven samples, may increase statistical power and thus provide a stronger test of cultural heterogeneity than a haphazard self-selected sample. For example, the current distribution of samples and participants (see Figure *1*) suggests that the researchers could have greatly reduced the participants at US sample sites and more evenly distributed sites between non-US Majority World countries without compromising on statistical power – indeed, this approach might have ironically boosted statistical power to detect heterogeneity across sites by reducing the degree to which USA data overwhelmed whatever signal of heterogeneity was present. Put differently, fewer but better samples could increase power. Whatever the details, the current approach of sample-sites self-selecting into a multi-site approach not only invites selection bias but also lacks a theoretical prediction of how cultural differences matter.

Drawing on simulations to determine the required sample size

During the design stage, simulations can help to better understand the required sample size given the expected effect of cultural differences. We provide the code to a simulation that models the characteristics of the ML2 environment, or more broadly, of a multi-site project investigating how culture moderates behavior (see Supplementary Section 1 for a full description of the simulation). Future studies can adapt this simulation to investigate which design features of cross-population multi-site studies (such as future Many Labs studies) would have the greatest power to detect cross-cultural heterogeneity in effects. As far as we can tell, ML2 did not run power simulations, which may have led to an overall lack of statistical power to detect all moderating effects of culture, even if they had been present. The lack of discussion of the power of moderation tests may suggest that the power of heterogeneity tests was simply assumed to be high because the overall power to detect main effects was high. We encourage researchers to more explicitly reflect on the power of statistical tests that drive key inferences, rather than focusing only on the main effect of statistical power.

Administering fewer studies per participant

The implementation of the replication featured a long list of studies presented at once (ML2 had 28 effects overall - 13 in Slate 1 and 15 in Slate 2). The effects were bundled and sequentially administered to participants in an extensive protocol online and in the lab. Despite being randomized in order, it is reasonable to assume that any context-specific task could produce noisy responses in such a setting. Exploring the effects of order provides only a partial solution, since different sequences may have different effects on different individuals and in different populations—resulting in greater measurement error. An effort to estimate the reproducibility of particular studies should endeavor to provide a strong test of those effects, by closely recreating the situation experienced by the original participants. Having participants participate in such a lengthy series of studies is at odds with how the original studies were conducted and are likely to result in them being less engaged with the studies, which would ultimately provide a weaker test of reproducibility. Future studies may thus consider administering fewer studies per participant.

Pre-registration cannot solve problems in theory and methodology

Pre-registrations can help improve scientific protocols and avoid statistical malpractices such as *p*-hacking (Nosek et al., 2018; Open Science Collaboration, 2015). But pre-registrations are just one approach in the toolbox to improve the quality of psychological science (Nelson et al., 2018; Nosek et al., 2022). One important aspect of a rigorous, replicable science is a good theory (Gervais, 2021; Muthukrishna & Henrich, 2019; Nosek et al., 2018), which has received less attention in the methodological reforms in psychological science. An inclusive and interdisciplinary framework, such as cultural evolutionary theory, can help to integrate theoretical approaches from different fields and help navigate the much-needed methodological changes in psychology (Gervais, 2021; Schimmelpfennig & Muthukrishna, 2023).

Operationalization of cultural difference

Measuring culture, and cultural differences for that matter is not easy. But whatever conceptualization of culture is chosen, future studies should ensure that cultural differences are conceptualized in a careful, theory-driven approach. The ML2 approach to decompose the WEIRD backronym may seem practically plausible but is not informed by a theoretical understanding of cultural variation. Other approaches to operationalizing measures of cultural differences via CF_{ST} (Muthukrishna et al., 2020), may provide better alternatives to capture population variability inductively. But a related important point is to avoid conflating cultural differences with cross-national differences. Some of the largest cultural differences are found when comparing state societies with non-state societies (Henrich et al., 2005), which ML2 did not sample from at all. Culture is not just cross-national, and not necessarily linear, but is embedded in intersecting distributions of cultural traits within societies (Cohen & Varnum, 2016; Muthukrishna & Henrich, 2019; Schimmelpfennig et al., 2021; Uchiyama et al., 2022), geographical regions (Talhelm et al., 2014), religious differences (White et al., 2021), exposure to markets (Enke, 2022; Henrich et al., 2005), social classes (Cohen & Varnum, 2016; Kraus et al., 2012), ethnicities (Desmet et al., 2017), kinship systems (Schulz et al., 2019) and political orientations (Desmet & Wacziarg, 2018; Ehret et al., 2022; Talhelm et al., 2015). Alternative approaches such as operationalizing cultural clusters by online behavior (Obradovich et al., 2022) should be considered.

WEIRD beyond samples

Lastly, biased participant samples are only one part of the WEIRD people problem. Many fields of psychological and behavioral science are also heavily biased towards WEIRD topics, WEIRD researchers, and WEIRD institutions, which can further bias the types of questions researchers ask of their WEIRD samples (for a further theoretical and methodological critique, see Gervais, (2021)).

CONCLUSION

Large-scale research efforts involving different research teams from around the globe are a critical part of advancing the field of psychological science in the future. They help address both the replication crisis (Open Science Collaboration, 2015), and the WEIRD people problem (Apicella et al., 2020; Henrich et al., 2010), but not necessarily the problem in theory (Muthukrishna & Henrich, 2019). We applaud the authors of the Many Labs 2 study (Klein et al., 2018), for their efforts to contribute to the field with an ambitious research project including participants living in 36 countries. It is our admiration that drives the critiques laid out in this paper that we hope will help motivate the theoretical and methodological changes needed to properly test the moderating role of culture.

Author contributions: ***Conceptualization:*** Robin Schimmelpfennig, Will Gervais, Ara Norenzayan, Steven Heine, Joseph Henrich, Michael Muthukrishna; ***Formal Analysis:*** Rachel Spicer, Robin Schimmelpfennig, Cindel White; ***Simulation:*** Will Gervais; ***Writing – Original Draft Preparation*:** Robin Schimmelpfennig, Rachel Spicer, Cindel White, Will Gervais, Ara Norenzayan, Steven Heine, Joseph Henrich, Michael Muthukrishna; ***Writing – Review & Editing*:** Robin Schimmelpfennig, Rachel Spicer, Cindel White, Will Gervais, Ara Norenzayan, Steven Heine, Joseph Henrich, Michael Muthukrishna.

Conflict of Interest: "The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article."

Acknowledgments: None.

Funding: M.M. acknowledges support from the Canadian Institute for Advanced Research (CIFAR) grant no. CP22-005 and Templeton World Charity Foundation grant TWCF0612. R.S. acknowledges support from the Swiss National Science Foundation (grant no. 100018_185417/1).

Pre-registration: https://osf.io/qrdxc/

Supplemental Material:

- OSF Folder containing Pre-registration, Simulation Files, and Supplementary Information: https://osf.io/qrdxc/
- The ML2 Paper: http://journals.sagepub.com/doi/10.1177/2515245918810225
- ML2 OSF Folder: https://osf.io/ux3eh/
- Data availability: The data we used for analysis is secondary data that was kindly shared with us by the authors of the Many Labs 2 study. It contains personally identifiable information of participants and thus cannot be publicly shared beyond what is available at ML2's repository (https://osf.io/ux3eh/). We have shared our code and the intermediate and final results of our analysis at https://osf.io/qrdxc/.

REFERENCES

Apicella, C. L., Norenzayan, A., & Henrich, J. (2020). Beyond WEIRD: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior*, *41*(5), 319–329. https://doi.org/10/ghndmz

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Borenstein, M., Hedges, L. V. P. T., Higgins, J., & Rothstein, H. (2011). *Introduction to Meta-Analysis*. John Wiley & Sons.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cohen, A. B., & Varnum, M. E. (2016). Beyond East vs. West: Social class, region, and religion as forms of culture. *Current Opinion in Psychology*, *8*, 5–9. https://doi.org/10.1016/j.copsyc.2015.09.006

Desmet, K., Ortuño-Ortín, I., & Wacziarg, R. (2017). Culture, Ethnicity, and Diversity. *American Economic Review*, *107*(9), 2479–2513. https://doi.org/10.1257/aer.20150243

Desmet, K., & Wacziarg, R. (2018). The Cultural Divide. *NBER Working Paper*, *w24630*, 53. https://doi.org/10/ghfjvm

Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C., & Vogt, S. (2022). Group identities can undermine social tipping after intervention. *Nature Human Behaviour*, *6*, 1–11. https://doi.org/10.1038/s41562-022-01440-5

Enke, B. (2022). Market exposure and human morality. *Nature Human Behaviour*, 1–8. https://doi.org/10.1038/s41562-022-01480-x

Gervais, W. M. (2021). Practical Methodological Reform Needs Good Theory. *Perspectives on Psychological Science*, *16*(4), 827–843. https://doi.org/10/ghxc54

Heine, S. J., & Norenzayan, A. (2006). Toward a Psychological Science for a Cultural Species. *Perspectives on Psychological Science*, *1*(3), 251–269. https://doi.org/10/cwtzfw

Henrich, J. (2020). *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. Farrar, Straus and Giroux.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795–815. https://doi.org/10/cp87gg

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A. M., Yilmaz, S., & Silk, J. B. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, *4*(1), 36–44. https://doi.org/10/gf9hmp

Huang, Y., Tse, C.-S., & Cho, K. W. (2014). Living in the north is not necessarily favorable: Different metaphoric associations between cardinal direction and valence in Hong Kong and in the United States: Metaphoric Association. *European Journal of Social Psychology*, *44*(4), 360–369. https://doi.org/10/gncnhq

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., & Puranen, B. (2014). *World Values Survey: All Rounds—Country-Pooled Datafile 1981-2014*. https://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp

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., Bialobrzeska, 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. https://doi.org/10.1177/2515245918810225

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, *63*(3), 190–194. JSTOR.

Kraus, M. W., Piff, P. K., Mendoza-Denton, R., Rheinschmidt, M. L., & Keltner, D. (2012). Social class, solipsism, and contextualism: How the rich are different from the poor. *Psychological Review*, *119*, 546–572. https://doi.org/10.1037/a0028756

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 1. https://doi.org/10.1038/s41562-016-0021

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychological Science*, 095679762091678. https://doi.org/10/ggxkjb

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Muthukrishna, M., Henrich, J., & Slingerland, E. (2021). Psychology as a Historical Science. *Annual Review of Psychology*, *72*(1), 717–749. https://doi.org/10/ghrnb6

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*, 291–310. https://doi.org/10.1037/0033-295X.108.2.291

Norenzayan, A. (2018). Some Reflections on the Many Labs 2 Replication of Norenzayan, Smith, Kim, and Nisbett's (2002) Study 2: Cultural Preferences for Formal Versus Intuitive Reasoning. *Advances in Methods and Practices in Psychological Science*, *1*(4), 499–500. https://doi.org/10/gkct26

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know. *Psychological Bulletin*, *131*, 763–784. https://doi.org/10/bgtcvt

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*(5), 653–684. https://doi.org/10/c4rc6h

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

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*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Obradovich, N., Özak, Ö., Martín, I., Ortuño-Ortín, I., Awad, E., Cebrián, M., Cuevas, R., Desmet, K., Rahwan, I., & Cuevas, Á. (2022). Expanding the measurement of culture with a sample of two billion humans. *Journal of The Royal Society Interface*, *19*(190), 20220085. https://doi.org/10.1098/rsif.2022.0085

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10/68c

Quintana, D. (2023). A guide for calculating study-level statistical power for meta-analyses. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31219/osf.io/js79t

Schimmelpfennig, R., & Muthukrishna, M. (2023). Cultural evolutionary behavioural science in public policy. *Behavioural Public Policy*.

Schimmelpfennig, R., Razek, L., Schnell, E., & Muthukrishna, M. (2021). Paradox of Diversity in the Collective Brain. *Philosophical Transactions of The Royal Society B*, *377*(1843). https://doi.org/10.1098/rstb.2020.0316

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, *366*(6466), eaau5141. https://doi.org/10/ggckxh

Stroebe, W. (2019). What Can We Learn from Many Labs Replications? *Basic and Applied Social Psychology*, *41*(2), 91–103. https://doi.org/10.1080/01973533.2019.1577736

Talhelm, T., Haidt, J., Oishi, S., Zhang, X., Miao, F. F., & Chen, S. (2015). Liberals Think More Analytically (More "WEIRD") Than Conservatives. *Personality and Social Psychology Bulletin*, *41*(2), 250–267. https://doi.org/10.1177/0146167214563672

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., & Kitayama, S. (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science*, *344*(6184), 603–608. https://doi.org/10/sp4

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2020). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist*. https://doi.org/10/gg4vbg

Uchiyama, R., Spicer, R., & Muthukrishna, M. (2022). Cultural Evolution of Genetic Heritability. *Behavioral and Brain Sciences*, *2022*, e152. https://doi.org/10/gkct6d

White, C. J. M., Muthukrishna, M., & Norenzayan, A. (2021). Cultural similarity among coreligionists within and between countries. *Proceedings of the National Academy of Sciences*, *118*(37), e2109650118. https://doi.org/10/gmv9pd