Good fit is weak evidence of replication:

Increasing rigor through prior predictive similarity checking

Wes Bonifay[1,2], Sonja D. Winter[1,2], Hanamori F. Skoblow[3], & Ashley L. Watts[4]

[1] Educational, School, & Counseling Psychology, University of Missouri

[2] Missouri Prevention Science Institute, University of Missouri

[3] Human Development & Family Science, University of Missouri

[4] Psychological Sciences, Vanderbilt University

This manuscript has been accepted for a forthcoming Special Issue of *Assessment* on "Assessment, Measurement, and Registered Replication." It has not yet been copy-edited.

Funding Statement: Bonifay and Skoblow were supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D210032. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Replication provides a confrontation of psychological theory, not only in experimental research, but also in model-based research. Goodness-of-fit (GOF) of the original model to the replication data is routinely provided as meaningful evidence of replication. We demonstrate, however, that GOF obscures important differences between the original and replication studies. As an alternative, we present Bayesian prior predictive similarity checking: a tool for rigorously evaluating the degree to which the data patterns and parameter estimates of a model replication study resemble those of the original study. We apply this method to original and replication data from the National Comorbidity Survey. Both data sets yielded excellent GOF, but the similarity checks often failed to support close or approximate empirical replication, especially when examining covariance patterns and indicator thresholds. We conclude with recommendations for applied research, including registered reports of model-based research, and provide extensive annotated R code to facilitate future applications of prior predictive similarity checking.

Keywords: replication, goodness-of-fit, model evaluation, Bayesian statistics, informative priors

Good fit is weak evidence of replication:

Increasing rigor through prior predictive similarity checking

Empirical scientific progress hinges on replicability, or the ability to duplicate the results of a previous study using the same procedures with new data (Bollen et al., 2015). Previous investigations of the replicability of studies and theories in psychological science have focused primarily on experimental effects (e.g., Klein et al, 2014; Open Science Collaboration, 2015; Youyou et al., 2023), but many subfields of psychology rely on statistical modeling rather than experimental design. Researchers in these areas, just like those in more experimental settings, must examine the degree to which their statistical models are replicable. As Neyman stated, "Models become plausible by repetition" (cited in Cudeck & Henly, 1991); that is, if statistical models are to be interpreted as accurate representations of some meaningful, generalizable phenomena, their statistical properties should be replicable.

One way to quantify model replication (whether in regression analysis, structural equation modeling, item response theory, network modeling, or any other statistical modeling framework) is with goodness-of-fit (GOF) of the model to the observed data. Over time, many psychological researchers have assumed that model replication is indexed primarily by reproducing the GOF from a previous study: "[the] improved model fit … replicate[d] the findings" of earlier research (Whiteman et al., 2022, p. 132), "the best fit … replicated previous findings" (Giuntoli et al., 2021, p. 1668), "substantially better fit … replicated the classical … approach" (Fernandez de la Cruz et al., 2018, p. 608), and so on. Paruzel-Cazchura and Blukacz (2021) even went so far as to characterize their GOF, not as evidence, but as *proof* of replication success: The "very good fit … proved the replicability of the overall structure" (p. 16). Indeed, the practice of relying exclusively on good fit as an index of model and theory replication has

buttressed, even generated, entire bodies of research (e.g., structural models of psychopathology; Kotov et al., 2021). This practice warrants careful consideration, as its ramifications bear important implications for scientific progress in psychology. Here, we argue that merely replicating the good fit of a model is insufficient support for the original statistical model and its underlying theory.

Consider the simulated covariance matrices shown in the upper row of Figure 1. The matrix on the left (panel B) represents the covariances among the variables in an original study, and the other matrices (panels C and D) represent the covariances among those same variables across two replication data sets. Thus, this figure represents the typical model replication scenario, in which the same structure is fit to different data sets and the parameters are allowed to freely vary. Despite the obvious differences in these data patterns, a model with two correlated factors fit each covariance matrix well (specifically, their Comparative Fit Indices [CFI] were high, i.e., ≥ .95). Strong GOF indicates that the model sufficiently represents the covariances within the original data and both replications, but it does not indicate whether each model reflects the same relationships among the variables of interest. As evidenced by Figure 1, a researcher relying on GOF alone risks missing potentially meaningful differences in these data patterns. To further complicate matters, the bottom row of Figure 1 shows that fitting the same model to each data matrix resulted in wide-ranging parameter estimates. For example, λ_{21}, the (standardized) factor loading of the second indicator on the first factor, was estimated at .90, .57, and -.63 for the three data matrices, and ψ_{21}, the correlation between factors, ranged from near independence (.09) to almost total overlap (.96). All told, GOF tells us nothing about the degree to which the replication data set or its model parameters resemble those of the original study. A strong similarity between these components is crucial to evaluating replication success.

Goodness-of-fit and Replication. In their critique of model-data fit as a general theory-testing tool, Roberts and Pashler (2000) argued[1], with support from philosophy of science and the history of psychology, that "showing that a theory fits data … is nearly meaningless" (p. 361; see also Vanpaemel, 2020). In particular, they identified three aspects of GOF that preclude it from providing strong theoretical support. Namely, good fit:

(1) Does not clarify what a theory predicts;

(2) Does not clarify the variability of the data; and

(3) Does not consider the a priori likelihood that the theory will fit any plausible data.

Accordingly, Roberts and Pashler reasoned that GOF only yields meaningful support for a theory when both data and theory are constrained, that is, when the data are not too variable and the theory is not too flexible. In a single study, such constraints may be difficult to formulate and impose on the data and/or model, in part due to a lack of reference criteria (e.g., *What does it mean to say that data are "not too variable"? Variable from what? And to what degree?*).

In replication efforts, however, there is a clear referent for characterizing the variability of the data and model parameter estimates: the original study. Thus, we can naturally extend Roberts and Pashler's (2000) three arguments to replication:

(1) Regarding the original study that the researcher intends to replicate, GOF does not predict anything of substantive value about the replication outcome. That the model fit well in the original study has no bearing on the replication of more important inferential aspects of statistical modeling, such as data patterns and parameter estimates.

[1] Although Roberts and Pashler (2000) focused on testing psychological *theories*, statistical *models* are representations of such theories (e.g., Fried, 2020; Scheiner & Holt, 2019; van Rooij, 2022). Thus, their critique of GOF in the context of theory testing also applies to the use of GOF in statistical model evaluation.

(2) GOF reveals nothing about the similarity of the original and replication data. Two data sets may be characterized by markedly distinct data patterns, potentially reflecting different data-generating mechanisms, but the model may obscure any meaningful data dissimilarities and thereby adversely impact the accuracy of inferences.

(3) If a model has a high degree of fitting propensity (i.e., an inherent tendency to fit well; Preacher, 2006; see also Bonifay & Cai, 2017; Falk & Muthukrishna, 2023), then good fit to the original and replication data sets will be unsurprising and of minimal scientific value. That is, when using a model that is predisposed to fit well, GOF may replicate regardless of the particular data patterns that researchers are intending the model to represent.

Ideally, researchers who establish that the replication data are no more variable than the original data, and that the parameter estimates in the replication data set reflect no more flexibility than those in the original data set, can be highly confident that their findings provide meaningful support for the theory underlying the statistical model.

Setting the Replication Target. In psychology, perfect similarity between studies is impractical and likely unnecessary, in part due to intrinsic population heterogeneity (McShane et al., 2019). Instead, researchers should focus their aim on the particular aspects of the original study that they wish to replicate. To that end, Figure 2 displays more reasonable target-setting in replication studies. The outer circle of the target depicts the status quo in the social sciences, as evidenced by the quotes referenced earlier: Establishing that the original model achieved good fit to the replication data, regardless of any empirical characteristics of the original study. As detailed, this practice offers the weakest support for the original theory.

Having established fit replication as the easiest and least beneficial target, we can now

focus our aim on more relevant (albeit more challenging) regions of the target in Figure 2. A researcher who wants to investigate the broader theoretical implications of the original study can aim at the *underlying theory* by specifying hypotheses regarding data or parameter estimates (e.g., "To support the underlying theory about the positive association between x and y, replication coefficient β^* must be any positive value"). Theory-informed replication improves upon fit replication and is an important target deserving of its own study. Accordingly, we reserve this topic for future research.

More ambitiously, a researcher who wants to directly replicate earlier empirical findings can aim at the original study. This endeavor can involve statistical tests of *approximate* or *close* empirical replication of both the original data and the original model (the terms *approximate* and *close* are subjective and allow for researcher discretion, but must be defined transparently, as we discuss later). Regarding the data, the researcher could inspect, for example, whether the covariances within the replication data are approximately or closely mirroring the covariances among the original variables. Regarding the model, the researcher could test whether the replicated parameter estimates are fairly similar to the original estimates (e.g., "In the original study, β_1 = .55; for approximate replication, β_1^* must be between .4 and .7") or exceedingly similar (e.g., "… for close replication, β^* must be between .52 and .58"). Passing such tests will provide more compelling evidence that the model has captured the same meaningful signal in both the original and replication studies.

Riskier Tests of Model Replication. In addition to helping researchers to focus their aims, the target displayed in Figure 2 is also useful in that it presents a series of increasingly stringent tests. Roberts and Pashler (2002) summarized the shortcomings of GOF by declaring: "When used to evaluate complex psychological theories, goodness-of-fit tests have been too easy to pass" (p.

605). The same conclusion applies to model replication: Researchers who represent their theories as statistical models should deemphasize GOF and instead rely on more challenging tests of replication success. Fit replication is the easiest target to hit (and for certain models, may entail essentially zero risk of failure; Bonifay & Cai, 2017) and, therefore, provides the weakest support for the original findings.

At each inner ring of the target, the risk of failure becomes greater, but the reward for success becomes ever stronger evidence of meaningful replication. Theory-informed replication is more stringent than fit replication, and the reward is some evidence that the replication supports the theory underlying the original study, if not the original findings. Approximate empirical replication is much riskier, and the payoff is strong evidence that the replication data and/or parameter estimates resemble those of the original study. Close empirical replication is the riskiest test: Researchers are unlikely to clear this high bar, but doing so would provide exceedingly strong evidence that the replication is, in fact, a reproduction of the original findings. In sum, every consecutive ring of this target presents a more challenging test of replication, and "risky tests are the most efficient means of gauging a theory's mettle" (Waller & Meehl, 2002, p. 331; see also Mayo, 1997).

In the following, we present a statistical method for rigorously quantifying the similarity between original and replicated data and parameter estimates. We then consider a real-world example in the context of modeling the latent structure of psychopathology. Finally, we provide recommendations for future replication research, with specific guidelines for preregistrations and registered reports of replication studies involving statistical models. Our method and open code will allow researchers to conduct this sequence of riskier tests, and thereby gain deeper insights into the strength of their evidence for replication success.

Methods

An Assumption. Before elaborating further, we offer a disclaimer: Replication studies and our methods for evaluating them are only valuable to the extent that the original work is worth replicating (Devezer et al., 2021). One component of good science (though not the only one; Haig, 2022) is the repeatability of previous research. But replicating work that was based on a poor theory, improper study design, and/or methodological deficiency is a fool's errand. If the original study was weak, then a successful replication effort–even one conducted through rigorous methods like those we promote here–will simply reproduce weak results (Patil et al., 2016).

In the context of statistical model replication, a critical concern is that the model is derived from a strong theory about the phenomena under investigation, where "strong" denotes a precise and unambiguous postulation regarding the relationships among the variables of interest (cf. Fried, 2020). Beyond a strong theoretical basis and corresponding statistical model, an original study that is worthy of replication should also entail a sufficiently large and representative sample (Anderson & Kelley, 2022), clearly operationalized constructs assessed with well-evaluated measures (Ioannidis, 2005), and transparent decision-making (Nosek et al., 2022). Simply put, the methods described hereafter assume that the original study is one that warrants a replication effort.

Prior Predictive Model Checking.

To formally investigate the similarity between the original and replication data and parameter estimates, we use Bayesian prior predictive model checking (PrPMC; Box 1980,

Evans & Moshonnov, 2006, Gelman et al., 2017).[2] In a Bayesian analysis, we are generally interested in the posterior distribution, $p(\theta|y)$, which we can compute by applying Bayes' theorem to some known y and unknown θ:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta). \quad (1)$$

Here, y is an observable variable and θ is an unknown parameter, and both can be vector-valued (i.e., can represent more than one variable or parameter). Per Bayes' theorem, $p(\theta)$ is the prior distribution for the parameter and $p(y|\theta)$ is the likelihood function for the parameter (given the data). The inclusion of the prior in the model allows us to evaluate the implications of our sampling model (i.e., the likelihood function) *before* including any observed data into the analysis. PrPMC allows us to evaluate such implications.

Briefly, PrPMC consists of generating predictive samples for each observed variable in our model, based solely on the prior distributions placed on the model parameters. These predictive samples represent hypothetical observed samples that are plausible under the expectations embedded in the prior distributions. Say we are interested in estimating the average height of adult rock climbers in the United States (US) and we expect that rock climbers are slightly below or above the US average of 168 cm (5'6"; Fryar et al., 2021). We quantify this expectation–-and our uncertainty about it–-by specifying $p(\theta) \sim N(\mu = 168,\ \sigma = 10)$: a Normal distribution with hyperparameters denoting a mean of 168 cm and a standard deviation of 10 cm. From this distribution, we can generate *prior predictive* samples of heights. Individual samples might have a mean as low as 138 cm (4'5'') or as high as 198 cm (6'5''), but across

[2] For further details on Bayesian inference, please see our OSF site (https://osf.io/q6rvf/), where we have uploaded a list of readings on introductory Bayesian statistics, Bayesian SEM, specification of suitable priors, and prior predictive model checking, along with links to online applications and other code examples that illustrate different components of prior specification.

samples, the average will be 168 cm. Based on our knowledge of adult rock climbers, we can use this range to change our prior distribution *before* we collect or analyze any data (e.g., by increasing μ and/or decreasing σ to reduce the plausibility of generating samples with a mean height as low as 138 cm).

The previous example demonstrates the typical use of PrPMC, to refine a model (in terms of the prior and sampling model) before any data are collected. But we can also use PrPMC to evaluate our model in light of the observed data (e.g., Zundert et al., 2022). This extension of PrPMC is akin to the more familiar techniques of *posterior* predictive model checking (PPMC; Guttman, 1967; Rubin, 1984). Both focus on some feature of the observed data and its discrepancy from that same feature across all predictive samples. This feature can be any statistical property of the data itself or any statistical result that can be obtained by fitting a model to data (Gelman et al., 1996). When this feature depends on the observed data alone, it is called a *test statistic* (e.g., median, quantiles, range). Using our rock climber example, setting the mean as a test statistic would entail comparing the mean height in an observed sample of rock climbers to the mean heights in all predictive samples. When this feature depends on both data and model, it is called a *test quantity* (e.g., GOF indices, parameter estimates; Gelman et al., 2013). A regression model of rock-climbing speed might include height as a predictor; the coefficient associated with height would be a test quantity computed by fitting the regression model to the observed sample and to every predictive sample.

More formally, the test statistic or quantity T in the observed sample y is compared to the same test statistic or quantity obtained from the simulated predictive samples y^{pred} (see the Appendix for technical details). For PrPMC, a *prior predictive* p-*value* (*prpp*) is then computed to quantify the likelihood of T in the distribution of the prior predictive samples:

$$prpp = p\left(T(y) \geq T\left(y^{\text{pred}}\right)\right). \quad (2)$$

Values near the extremes of this predictive distribution (e.g., *prpp* ≤ .05 or ≥ .95) indicate the presence of systematic differences between the observed sample and the prior predictive samples. Values near .50 indicate that the observed *T* is in near the center of the predictive distribution, meaning it aligns with the *T* we would expect to see, based on our priors. In our rock climber example, a *prpp*-value of .99 for *T*(mean height) would reflect that the climbers in the observed sample were at the upper extreme of the predictive distribution, meaning they were much taller, on average, than we anticipated. Conversely, a *prpp* = .50 would indicate that the observed heights are exactly what we predicted.

Prior Predictive Similarity Checking. In the context of model replication, we propose a novel application of PrPMC that we refer to as *prior predictive similarity checking*. This process will allow researchers to quantify the degree of similarity between the original and replication data, via test statistics, and between the original and replication model parameter estimates, via test quantities.

For the data similarity check, researchers can inspect *prior-data disagreement* (Evans & Jang, 2010; Evans & Moshonov, 2006), which, in the context of replication research, concerns the conflict between our prior expectations (as informed by the original study) and the observed replication data. A *prpp*-value between .05 and .95 would support that the original and replication data are similar in terms of the chosen test statistic *T* (e.g., interitem correlations), whereas *prpp* < .05 or > .95 would indicate dissimilarity. Passing this check would strengthen claims of replication success by ensuring that the data patterns do not resemble the disparate matrices at the top of Figure 1. By directly checking the data patterns, this method is conceptually aligned with specification of a *data prior*, which has been recently recommended as

a tool for stronger theory testing (Vanpaemel, 2020; Villarreal et al., 2023). In the present context, prior specification greatly benefits from reliance on the original findings.

For the parameter estimate similarity check, researchers must first fit the selected model to the prior predictive data and the replication data and then evaluate relevant test quantities. Typically, a test quantity is computed for each draw from the posterior, meaning both the observed data and the model are involved in determining its value. However, we propose that test quantities, particularly those relating to parameter estimates, can also be computed through PrPMC. Specifically, we propose that draws from the prior pushforward distribution can be used to compute a test quantity about the model parameters (e.g., factor loadings in each prior pushforward sample). The distribution of those test quantities can then be compared to that same test quantity obtained by analyzing the observed data through traditional estimation methods (e.g., maximum likelihood). This approach allows us to directly test our expectations about replicated parameter values.[3]

Returning to Figure 2, note that the series of prior distributions at the bottom of this figure correspond to the rings of the replication target. The outer ring is associated with a *diffuse* (i.e., flat) prior distribution of widely varying data characteristics and/or parameter estimates, resulting in a relatively risk-free and mostly worthless check of similar GOF. Whereas the bullseye reflects a model with a highly *informative* (i.e., peaked) prior distribution of narrowly constrained data characteristics and/or parameter estimates, resulting in an appropriately stringent check of key components of the original and replication studies. In either case, if the

[3] Zondervan-Zwijnenburg (2019) also proposed a method for testing replication of parameters using PrPMC, though it requires a) access to the data from the original study, and b) specification of one or more replication hypotheses that rely on (in)equality constraints for sets of parameters. Our method is more versatile as it does not require the original data or specification of parameter constraints (though could accommodate the latter).

researcher finds *prpp*-values between .05 and .95, then they will have evidence that the similarity check, at the preferred level of risk, has been passed.

Why Bayes? In psychological modeling and assessment research, the Bayesian statistical framework may be less familiar than the traditional frequentist approach. We contend that it offers three key advantages in the context of model replication efforts. First, in frequentist inference, the point estimate of a statistical parameter is typically prioritized over its uncertainty, whereas the primary aim of Bayesian inference is to characterize that uncertainty (Kruschke et al., 2012). In the context of factor loading replication, for example, a traditional frequentist analysis would compare the replication loading to the point estimate (e.g., $\hat{\lambda}_{ij} = .60$) of the corresponding original loading. Doing so assumes that the original point estimate represents the true factor-indicator association in the initial study. By contrast, a Bayesian prior predictive similarity check would compare the replication loading to a *distribution* of theoretically plausible loadings that better capture our uncertainty about the true factor-indicator association in the original study. The Bayesian emphasis on uncertainty also more closely reflects our own view of replication studies in general. A single original study has a high degree of uncertainty; this initial uncertainty can be reduced (to varying degrees) by a successful replication of key features of the data and/or model parameters, or increased (to varying degrees) by a failed replication of such features.

Second, though frequentist standard errors represent uncertainty in point estimates, their magnitude may not reflect a reasonable level of similarity. In fact, in the application example below, we initially attempted to base the priors for close replication on the maximum likelihood estimated standard errors, which were exceedingly small due to the large sample size of the original empirical data. The resulting target for close replication became so narrow as to be

virtually impossible to hit. Conversely, an original study with a meager sample size may have large standard errors, and the corresponding replication target will be too easy to hit. Given this limitation of frequentist standard errors, we argue that the Bayesian framework offers more control over setting the replication target(s).

Third, although frequentism provides an analog to Bayesian model checking in the form of bootstrapping (Efron & Tibshirani, 1993; Rubin, 1981), that procedure is not flexible enough for rigorous testing of model replication. Bootstrapping is limited by its strict reliance on the observed data, and thus entails the bold (and often indefensible) assumption that the observed data are a worthy source from which to resample. In contrast, PrPMC does not require that any data have been collected, but rather simulates a distribution of what data could look like, given prior expectations (as in our earlier rock-climbing example). So while frequentist bootstrapping and Bayesian model checking are conceptually similar, they differ in an important way: Bootstrapping compares the replication data to the original data[4], whereas our Bayesian approach compares the replication data and/or model parameters to the hypothetical data and/or model parameters that one would expect to see in a successful replication (with flexibility regarding the target for success).

In sum, we forgo more familiar frequentist methods because a Bayesian perspective allows us to emphasize uncertainty in the original study estimates, maximize control over the replication target, and eschew the original observed data. In our view, these facets are critical for thorough model replication testing and readily obtainable within a Bayesian framework. In what

[4] One might argue that the parametric variant of bootstrapping (Efron, 1979)—wherein one resamples from the observed data likelihood rather than the data itself—could be equivalent to Bayesian model checking. However, the parametric bootstrap does not allow for user-specification of the informativeness of the distribution, which is crucial for consideration of different replication targets (Figure 2).

follows, we provide a worked application of our method (with corresponding R code and other resources available at https://osf.io/q6rvf/) to increase familiarity and comfort with the method we propose.

Empirical Application

To demonstrate prior predictive similarity checking, we will use two data sets, one "original" study and one "replication," to demonstrate the applicability of our proposed methods: The National Comorbidity Survey (NCS; Kessler et al., 1994) and the National Comorbidity Survey Replication (NCS-R; Kessler & Merikangas, 2004), respectively. NCS is a national probability sample of adults aged 15 to 61 (N = 8,098; M_{age} = 34, 53% female) collected between 1990 and 1992, and NCS-R is an independent sample of adults aged 18 to 99 (N = 9,282; M_{age} = 45, 55% female) collected between 2001 and 2003. Recruitment and consent procedures for NCS and NCS-R were approved by the Human Subjects Committees of Harvard Medical School and the University of Michigan.

We selected these data sets for three reasons. First, these data sets have been used in a wide array of well-cited studies on the latent structure of psychopathology (e.g., Krueger, 1999). Second, they are well-suited as "original" and "replication" data sets because their respective studies followed similar recruitment and sampling strategies, and comparable methods for assessing psychopathology. Third, both data sets are publicly available, allowing for full reproducibility of our analyses (NCS: https://www.icpsr.umich.edu/web/ICPSR/studies/6693, NCS-R: https://www.icpsr.umich.edu/web/ICPSR/studies/20240). Our project's Open Science Framework (OSF) page (https://osf.io/q6rvf/) contains a data dictionary of all indicators used in the analysis.

Psychopathology Assessment and Statistical Model. NCS and NCS-R assessed

psychiatric diagnoses using trained, nonclinician interviewers who administered a modified version of the World Health Organization Composite International Diagnostic Interview (for reliability, see Kessler & Üstün, 2004; Wittchen, 1994). As with previous studies (e.g., Krueger, 1999), we used binary (yes/no) indicators of lifetime diagnoses of mental disorders that were generated without hierarchical exclusion rules (with the exception of major depression, see our later description), which is standard in psychiatric epidemiology (Kessler et al., 1994). Diagnoses were based on the Diagnostic and Statistical Manual of Mental Disorders, version III (American Psychiatric Association, 1980) for NCS, and version IV (American Psychiatric Association, 1994) for NCS-R.

As mentioned, Krueger (1999) previously examined the structure of psychopathology in NCS. To the extent possible, our model adhered to the indicators used by Krueger (1999), with some minor deviations to maintain compatibility across the original and replication data sets. Krueger (1999) extracted a three correlated factors model with externalizing (tendencies toward poor behavioral and emotional control), distress (tendencies toward anxious misery) and fear (tendencies towards phobic responses to stressors) dimensions. In our model, externalizing was composed of alcohol dependence, drug dependence, and conduct disorder; distress was composed of major depression, dysthymia, and generalized anxiety disorder; and fear was composed of agoraphobia, panic disorder, social anxiety disorder, and specific phobia. In contrast with Krueger (1999), we used conduct disorder diagnoses as opposed to antisocial personality disorder diagnoses because NCS-R only makes the former available. We also relied on major depression diagnoses with hierarchical exclusion rules because NCS-R does not make depression diagnoses without hierarchical exclusion rules available.

Analysis. We use the *lavaan* R package (Rosseel, 2012) to estimate the three-factor

confirmatory factor analysis (CFA) model[5] described earlier, with mean- and variance-adjusted weighted least squares (WLSMV) estimation and theta parameterization. To start, we assessed GOF using the χ^2 test and the familiar benchmarks for the comparative fit index (CFI > .95), Tucker-Lewis index (TLI > .95), and root mean square error of approximation (RMSEA < .06), along with its confidence interval (CI) (Hu & Bentler, 1999). The parameter estimates and standard errors of these analyses formed the basis for the prior specifications that we used to test different degrees of similarity.

We then used the *blavaan* R package (Merkle & Rosseel, 2018) for similarity checking. To specify prior distributions for all loadings, we scaled each factor by fixing its variance to 1.0. This scaling does not affect thresholds, so threshold priors were based on the unstandardized estimates from the original study. For the associations between factors, we followed *blavaan* defaults by assuming that the hyperparameter values reflect correlations rather than covariances. In total, we compared three sets of prior specifications, which we mapped onto different rings of the replication target. Example specifications for factor loadings, thresholds, and factor correlations are shown in Figure 3, and Table 1 includes a detailed overview of these specifications. [6]

- *Close Replication:* In the most constrained condition, factor loadings and thresholds were assigned Normal prior distributions with mean hyperparameters set to the corresponding

[5] Herein, we demonstrate prior predictive similarity checking in the context of factor analysis, but the same methodology applies to other models. Readers who wish to extend this method to other models should consult our reading list (https://osf.io/q6rvf/) for relevant tutorials on prior specification and prior predictive model checking (e.g., van Zundert et al., 2022; Winter & Depaoli, 2023; Zondervan-Zwijnenburg et al., 2017).

[6] We focused on prior specifications that are currently available in accessible software. Future research might focus on multivariate prior distributions placed on the complete factor loading matrix, or priors placed on the observed item covariance matrix to facilitate comparison to a saturated model.

standardized factor loading/unstandardized threshold estimate and standard deviations set to 0.05. Factor correlations were assigned Beta prior distributions with mode equal to the estimated correlations from the original study. In addition, the hyperparameters were specified so that a narrow range around the mode was assigned the highest prior probability.

- *Approximate Replication:* Priors for the factor loadings and thresholds were the same as in the close replication condition, except that the standard deviation was increased to 0.50. Factor correlations were assigned Beta prior distributions with their mode equal (within .0001) to the original correlation estimates. In addition, the hyperparameters were specified so that most of the prior density covered values > 0, reflecting an expectation that the correlations should be positive.
- *Fit Replication:* In the least constrained condition, factor loadings and thresholds were assigned a Normal prior with a mean of 0 and standard deviation of 5 (for loadings) and 3 (for thresholds). Factor correlations were assigned a joint Lewandowski-Kurowicka-Joe (LKJ) prior with its eta hyperparameter set to 1 (in line with the *blavaan* default). Note that the label "fit replication" reflects the target, not the outcome.

Of course, different prior distributions could be specified, depending on how different researchers define "close" and "approximate." For example, one might decide that, for a close replication, the factor loading estimates in the replication study should be within .05 of the original factor loading estimate, which would be specified by setting the Normal standard deviation hyperparameter at .025, whereas another researcher might specify different criteria. Note that our OSF site (https://osf.io/q6rvf/) includes a list of selected readings on Bayesian modeling and prior specification, links to interactive prior specification examples, and annotated code to allow researchers to explore their own prior specifications.

Finally, to investigate data similarity at each level of risk, we considered a set of test statistics comprising all 45 pairwise correlations among the indicators. We chose to examine the correlations because in CFA, the model is assumed to represent patterns of association in the data; our analysis will allow us to determine the degree to which these patterns are similar across the original and replication data.[7]

To investigate similarity of parameter estimates, we considered four sets of test quantities. We selected three of these quantities because they comprise the fundamental components of CFA model interpretation: factor loadings, thresholds, and factor correlations. In our view, all CFA model replication efforts should inspect the degree to which these key parameter estimates are similar across the original and replication studies.

We also calculated the intraclass correlation coefficient (ICC) [specifically the ICC(3,1) variant; Shrout & Fleiss, 1979] to compare the replication loading *pattern* against each of the 2,000 loading patterns drawn from the prior pushforward distribution. In doing so, we followed Lane et al. (2016), who used ICCs to compare alcohol use disorder symptom profiles across multiple studies. There, they argued that if such patterns are dissimilar, then "it would not be reasonable to conclude replicability, except in a superficial sense" (p. 1770). In our context, ICC $\leq .00$ indicates a total lack of consistency between the actual replication loading pattern and the pattern one would expect, given the original results and the desired replication target, whereas ICC = 1.00 indicates perfect consistency. To aid in interpretation, Koo and Li (2016) discretized this ICC continuum into subranges for poor (ICC < .50), moderate ($.50 \leq$ ICC < .75), good ($.75 \leq$ ICC < .90), and excellent (ICC $\geq .90$) consistency. Our analysis considered loading pattern

[7] Part B of the Appendix includes a second set of test statistics (indicator response proportions), which we included to demonstrate that this method allows a researcher to explore similarity with respect to any feature of the data.

similarity across all ten psychopathology indicators and within each of the three factors.

In the least severe test of pattern similarity, loadings that are not constrained by an informative prior distribution (i.e., when targeting fit replication) will be free to take on a wide range of expected patterns. Any similarity between the observed replication pattern and these widely varying expectations will be, essentially, a chance occurrence. Thus, because the ICCs index consistency between the replication pattern and each of the 2,000 patterns from the prior pushforward distribution, we anticipate many ICCs near zero when targeting fit replication. In the most severe test, loadings that are constrained by a highly informative prior distribution (i.e., when targeting close empirical replication) will yield expected patterns that closely resemble the original loading pattern. The resulting ICCs will be consistently large if, and only if, the observed replication pattern actually conforms to this narrow range of expected patterns; but if the replication pattern diverges from these expectations, then most ICCs will be small, indicating a failure of this severe test.

Results

The three-factor CFA fit well to the original (NCS) data ($\chi^2[32] = 160.19$, $p < .001$, CFI = .982, TLI = .975, RMSEA = .022, 90% CI [.019, .026]), yielding the parameter estimates in Table 1. The same model also fit well to the replication (NCS-R) data ($\chi^2[32] = 124.70$, $p < .001$, CFI = .991, TLI = .987, RMSEA = .018, 90% CI [.014, .021]). Thus, the replication study effectively reproduced the original fit; but as we have demonstrated, a phrase like "reproduced the original fit" should be treated as a feel-good bromide rather than a scientific pronouncement. We turn now to investigating the similarity of the NCS and NCS-R data and parameter estimates.

Similarity of Data. To assess the similarity of the original and replication data, the first set of test statistics we considered were pairwise tetrachoric indicator correlations. Figure 4

presents the *prpp*-values from this initial data similarity check. When targeting close empirical replication via the priors in Table 1, the similarity check was an almost total failure: With the exception of two item pairs – indicators 3 (conduct disorder) and 4 (major depression) and indicators 2 (drug dependence) and 6 (generalized anxiety) – the NCS-R (replication) data were more correlated than we expected, given the NCS (original) data patterns. When targeting approximate empirical replication, the similarity check was mixed: 23 of the 45 item pairs (51.11%) were aligned with expectations. But when targeting fit replication (i.e., ignoring the original data patterns altogether), 100% of the item pairs passed the similarity check. In other words, a researcher who only cares about replicating fit will be able to hit that safe and easy target despite varying data patterns between studies (and potentially, in some cases, different data-generating mechanisms).

Figure 5 further elaborates on data (dis)similarity. The inset depicts all three prior predictive similarity checks of the correlation between indicators 9 (specific phobia) and 10 (panic disorder). The horizontal line denotes $T(y)$, the estimated correlation between indicators 9 and 10 in the replication data ($\hat{r}_{9,10} = .49$), and the bars represent the 90% interval of $T\left(y^{\text{pred}}\right)$, the prior predictive distributions of $r_{9,10}$ as implied by the original data. When the target was close empirical replication, $\hat{r}_{9,10}$ failed the similarity check, which could only be passed when $\hat{r}_{9,10} = (.26, .42)$. When the target was approximate empirical replication, $\hat{r}_{9,10}$ passed the similarity check because it fell within the range $(-.15, .61)$. When the target was merely fit replication (i.e., irrespective of the original data patterns), $\hat{r}_{9,10}$ also passed the similarity check. However, this third test was so risk-free as to be nearly impossible to fail: Any correlation within the range $(-.97, .97)$ would have passed. In fact, across all 45 bivariate correlations shown in Figure 5, the fit replication bar is always so large as to be virtually unmissable (in parallel to the

outer ring of the target in Figure 1). By contrast, the truncated size of the approximate replication bar denotes a much riskier prospect, and the close replication bar is so small that failure of the similarity check is near inevitable. Again, if these riskier checks can be passed, then claims of replication will be considerably stronger.

***Similarity of Parameter Estimates*.** Figure 6 presents the prior predictive similarity checks of the individual factor loadings. In each plot, the vertical line indicates the loading estimate from a traditional (frequentist) analysis of the replication data [$T(y)$], which is superimposed atop the prior pushforward distribution of the loading [$T\left(y^{\text{pred}}\right)$] for each replication target. Close empirical loading replication was evident for all indicators except 1 (alcohol dependence) and 7 (agoraphobia): Loadings of both indicators were higher than expected in the replication data. For approximate empirical replication and fit replication, all ten loadings passed the similarity check.

However, Figure 7A[8] shows that similarity between the original and replicated loading *patterns* was rarely achieved across all ten indicators. None of the patterns from the prior pushforward distribution reached good or excellent consistency with the observed replication loadings. When the target was close empirical replication, 12.05% of the ICCs exhibited moderate pattern consistency, but this outcome was much less likely when targeting approximate empirical replication, wherein only 1.40% of the ICCs exhibited moderate consistency. When targeting fit replication, only a single sample from the prior pushforward distribution (0.05%) exceeded the cutoff for moderate consistency (and just barely: ICC = .51).

Within each factor, pattern consistency varied considerably. Among the three externalizing indicators (Figure 7B), similarity was highly likely when targeting close empirical

[8] See also the table of ICC classification percentages at our OSF repository: https://osf.io/q6rvf/

replication, as 89.15% of all patterns showed moderate or better pattern consistency. This level of consistency far exceeded that of the approximate empirical replication (36.85%) and fit replication (5.85%) targets. The three distress indicators (Figure 7C) fared worse, as their ICCs indicated moderate or better pattern consistency in 40.00% of close empirical replications, compared to 20.90% of approximate empirical replications, and just 4.50% of fit replications. However, pattern consistency among the four fear indicators (Figure 7D) followed the opposite trend: When targeting close empirical replication, 100% of the distress loading patterns were extremely inconsistent (ICC M = .00, SD = .02). Yet, when targeting approximate empirical replication or fit replication, 4.80% and 1.27% of ICCs, respectively, showed moderate or better consistency. We provide an explanation for these seemingly counterintuitive results in our Discussion.

As for the thresholds (Figure 8), none of them passed the close empirical replication test. Thresholds in the replication data [$T(y)$] were far higher than expected [$T(y^{\text{pred}})$] for alcohol dependence, drug dependence, conduct disorder, dysthymia, and agoraphobia; and lower than expected for generalized anxiety disorder, specific phobia, and panic disorder. The approximate empirical similarity check also failed for alcohol dependence, drug dependence, dysthymia, and agoraphobia. When aiming at simple fit replication, the prior pushforward distributions were wide enough to accept every threshold as similar, no matter how much the estimates varied from the original findings.

Finally, we examined the similarity of the factor correlations, as shown in Figure 9. For all targets, the three correlations in the replication data were located within the central 90% of their corresponding prior pushforward distributions, with one exception: The replication correlation between the externalizing and fear factors was higher than expected.

Discussion

Jaynes (2003) wrote that "a false premise built into a model which is never questioned, cannot be removed by any amount of new data. Use of models which correctly represent the prior information that scientists have about the mechanism at work can prevent such folly in the future" (p. xvi). In model replication studies, the false premise is that GOF is sufficient evidence of replication. In this paper, we addressed the shortcomings of GOF by developing a method that correctly represents the prior information from the original study that one wishes to replicate. Prior predictive similarity checking enables researchers to rigorously compare–across a series of increasingly risky statistical tests–the data and/or parameter estimates of the original and replication studies. Importantly, although our empirical analysis involved one particular model, the general methodology can be extended to any replication study involving a model that is amenable to Bayesian prior specification and predictive model checking. We focused on CFA because this predominantly psychological modeling framework is likely familiar to a broad and diverse readership. The methods we have described can be directly extended to models that are less constrained (e.g., exploratory factor analysis) or more constrained (e.g., a one-factor model), as long as the exact same model is applied to both the original data and the replication data. In fact, and in contrast to measurement invariance testing and related techniques such as the alignment method (Asparouhov & Muthén, 2014), our similarity checking procedure is not tied to CFA or SEM at all, and can shed light on replication efforts involving any other model that can be estimated using Bayes' theorem.

Regarding our findings, it is important to underscore that the NCS and NCS-R data sets present a best-case scenario for psychological model replication. First, these data sets, rather than coming from replication attempts by disconnected parties, as is often the case in replication

research, were collected by the same core researchers and with consistent procedures. Second, the 3-factor CFA model that we considered has strong theoretical grounding in the psychopathology assessment literature. Third, the NCS and NCS-R both featured large sample sizes, which increase the precision of the data descriptives and model parameter estimates. Whenever original and replication data are large enough to effectively represent the population values, then close or approximate empirical replication is more likely. Thus, that some of the similarity checks were successful in this scenario is evidence of the importance of research design and data collection when targeting empirical replication; in typical modeling applications with inconsistent data collection, a weak underlying theory, and/or smaller samples, hitting the inner regions of the replication target will be an even more exacting test.

Consequently, our analysis of the NCS and NCS-R data uncovered several noteworthy patterns. First, although the three-factor CFA model fit exceptionally well to both data sets, the data patterns and parameter estimates of the original and replication studies often failed our checks of close and approximate empirical replication. In other words, fit statistics obscured key differences between the NCS and NCS-R studies and the relationships between participants' diagnoses and underlying psychopathology. Thus, a model that fits equally well to two data sets may be representing data from markedly different data-generating mechanisms, which raises the possibility that the latent factors differ in meaning across so-called replication data sets.

Second, although our analysis established similarity between the original and replication loadings in all but two instances (close empirical replication of the original alcohol dependence and agoraphobia loadings), similarity between loading patterns was more equivocal. Patterns among the fear indicators seemed particularly unusual, in that targeting close empirical replication resulted in total dissimilarity with all expected patterns. However, further

consideration reveals that the ICCs relate to two aspects of the similarity checks on individual loadings: their location (whether the replication loadings were consistently above or below the bulk of a given target distribution) and, less importantly, their accuracy (whether the replication loading hit the target, i.e., yielded a *prpp*-value between .05 and .95).

Consider, as an example, the close empirical replications of individual loadings in Figures 6 and loading patterns in Figure 7. For the externalizing factor, the replication loadings were consistently located above the peak of the close target distribution, though they missed the target for alcohol dependence. The corresponding ICC distribution in Figure 7B showed that the pattern of externalizing indicators in the replication study exhibited moderate-to-excellent alignment with the patterns one would expect when targeting close empirical replication.

For the distress factor, the replication loadings were not consistently located (major depression and generalized anxiety were slightly smaller than the modal expected loading, and dysthymia was considerably larger), but all three indicators hit the close empirical replication target. The corresponding ICCs in Figure 7C showed that the similarity between loading patterns was mixed: The replication loading pattern was inconsistent with 60.00% of the expected patterns, but realized moderate or better consistency with the other 40.00%.

For the fear factor, the replication loadings were inconsistently located above (agoraphobia and social anxiety) and below (specific phobia and panic disorder) the bulk of the close target distribution, *and* they failed to hit the agoraphobia target. The combined effect of these varying locations and imperfect accuracy resulted in the ICC distribution in Figure 7D, in which close empirical replication was impossible. Overall, close empirical replication of the original factor loading patterns appears to be more likely when the individual replication loadings that make up that pattern are consistently located and on target.

Third, we found that the threshold estimates were not similar between the original and replications. This finding has important implications for the generalizability of CFA models across critical subpopulations. With the binary indicators used here, the thresholds correspond to the prevalence of mental disorders in each particular data set. By definition, subpopulations with varied base rates of diagnoses–different genders, cultures, degrees of clinical severity, to name a few–will yield different thresholds in a CFA model. In turn, a CFA model that fits equally well across subpopulations may lead researchers to infer the same diagnosis, despite endorsement of different, even mutually exclusive, sets of criteria depending on the subpopulation (Boness et al., 2019; Lane et al., 2016).

Fourth, in general, it was easier to establish similarity of parameter estimates than similarity of data. The original and replication data patterns were decidedly dissimilar, especially when targeting close empirical replication. This finding recalls the literature on model complexity, defined by Pitt and Myung (2002) as "the property of a model that enables it to fit diverse patterns of data" (p. 422). The three-factor CFA model was able to fit the diverse patterns in the original and replication data, and a more flexible model would make close or approximate similarity an even more challenging target. In our view, this finding invites skepticism of the value of claims like "the model represents the data," particularly when the research task involves fitting one model to more than one data set, as in replication efforts. Therefore, even though our data similarity checks often failed, we maintain that researchers should attempt to detect and counteract model complexity by establishing that the data are not wildly dissimilar.

Implications of (dis)Similarity. Our methodology aligns with the recent definition of replication given by Nosek and Errington (2020): "Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research" (p. 2). Prior

predictive similarity checking offers a confrontation of theory rather than a confirmation of theory. Passing the similarity check does not imply that the statistical model, in either the original study or the replication, is "correct," nor does it provide definitive confirmation of theory. Similarity is not an arbiter of truth, but it does add support for the theory that was formalized in the statistical model. Even if the original study entailed strong theory, precise measurement, and no questionable research practices, a statistical model is an imperfect representation of the observed data and will always leave some lingering uncertainty about the generalizability of the findings. This uncertainty will be reduced by evidence of statistical similarity between the original and replication studies; the riskier the replication test, the greater the reduction in uncertainty (i.e., we can feel more certain that the findings generalize beyond the original context).

Likewise, failure to pass the similarity check does not imply that the attempted replication was flawed or invalid, nor does it prove that the original study or its motivating theory were incorrect. Rather, a lack of similarity increases, rather than reduces, the uncertainty about the generalizability of the original findings. Within our methodological approach, finding that key *prpp*-values are $\leq .05$ or $\geq .95$ does not refute the initial study or the replication, it simply makes us less convinced that the original findings generalize to other samples. In that sense, a failed similarity check is still a valuable scientific outcome.

Recommendations for Applied Research

Practical implications of our arguments surround the benefits of prior predictive similarity checking when planning and evaluating registered reports (Chambers, 2013; Lee et al., 2019; Simons et al., 2014). In a registered report, authors submit a preregistration plan for peer review; if the plan is accepted, the journal guarantees to publish the study results no matter the

outcome (i.e., even if the results are null). Despite the potential benefits of preregistration—mitigating questionable researcher practices and reducing the "file drawer problem" (Nosek et al., 2018; Rosenthal, 1979)—the process is not without its limitations. Researchers can preregister a weak hypothesis and assess it with a weak test, as occurs when claiming to replicate a latent factor structure by reproducing good fit alone. Thus, even if the enacted plan corresponds to the *a priori* plan, such a replication would instill little confidence in the results and its underlying theory. To that end, Vanpaemel (2019) stated that preregistration alone "does not guarantee risky predictions, strong tests, and solid evidence" (p. 219).

Our methods for setting a replication target and testing specific hypotheses strengthen the value of registered reports, in line with Vanpaemel's (2019) Really Risky Registered Report (though that template was designed for modeling in general rather than model replication). First, rather than using GOF as the sole evidence of replication, researchers should preregister their specific hypotheses, linked to the degree of target-setting involved in their replication. When setting the target, heed the advice of Dyson (1999): "it is better to be wrong than to be vague" (p. 48). Fit replication is an easier target to hit, but any conclusions about similarity of data or similarity of parameter estimates will be vague. Researchers who want strong evidence of replication should feel emboldened to include riskier tests in their registered reports. Although hitting the bullseye of close empirical replication is unlikely, demonstrating evidence of approximate empirical or nonempirical yet theory-informed replication will yield stronger support for the original study than can be garnered from fit replication alone.

Second, these plans should include the exact prior distributions that the researchers will use for similarity checking, like those we included in Table 1. For a close or approximate empirical replication, preregistration should address how well aligned the original and replication

elements must be to constitute a meaningful replication. Due to the subjectivity and arbitrariness of terms like “close” and “approximate,” we urge researchers to be precise and transparent about their operational definitions and the statistical formalization thereof (i.e., by reporting the exact probability distributions that correspond to their verbal definitions). For illustrative purposes, we set the prior standard deviations of the factor loadings at .05, .50, and 5.00 for close, approximate, and fit replication, respectively, but reasonable target-setting will likely vary across constructs and contexts. Our annotated code (https://osf.io/q6rvf/) allows researchers to modify our settings for use with their own data, model, and underlying theory. These settings should be thoroughly detailed in any registered report. Reviewers of such a registered report can also help the authors to specify appropriately risky hypotheses and priors.

Ultimately, any registered report should be transparent about the precise plan for evaluating model replication, beyond reliance on GOF. If, in the ensuing research, the observed data patterns and parameter estimates pass the similarity check, then the preregistered hypotheses will have passed a challenging test, and the reward will be far stronger claims of replication.

If, however, researchers are unable to pass a similarity check, they should detail all evidence of that failure and fully describe its implications for model replication and generalizability. For such an outcome, we recommend follow-up analyses aimed at exposing the source(s) of the replication failure. If the exact same measurement instruments were used in the original and replication studies, then researchers could conduct formal tests of measurement invariance, both between and within each sample, to illuminate any measurement bias and/or genuine substantive variation that could be attributed to population heterogeneity. More generally, researchers could examine the original and replication studies for any number of research design and analysis considerations (e.g., differences in data collection, adherence to

testing protocols, statistical software settings, treatment of missing values) that may have contributed to dissimilarity.

Conclusion

Replication efforts confront one's theory, not only in experimental settings, but also in applications of statistical modeling. Good fit to the replication data, which is often the sole evidence that researchers consider, has essentially no bearing on similarity to the original findings. To enhance current model replication practice, this paper presents the statistical method of prior predictive similarity checking. This approach is much riskier and more technically involved than fit assessment, and, in exchange, imparts a direct quantification of the degree to which important components of model replication resemble those of the original study. By adopting this method as a routine aspect of replication research and registered reporting, researchers will be equipped to evaluate their model replication efforts with rigor, precision, and conviction.

References

American Psychiatric Association. (1980). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.) (DSM-III). American Psychiatric Association, Washington DC.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.) (DSM-IV). American Psychiatric Association, Washington DC.

Anderson, S. F., & Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological methods*, 10.1037/met0000520. Advance online publication. https://doi.org/10.1037/met0000520

Asparouhov, T. & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). Reproducibility, replicability, and generalization in the social, behavioral, and economic sciences. Arlington, VA: National Science Foundation. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences

Boness, C. L., Lane, S. P., & Sher, K. J. (2019). Not all alcohol use disorder criteria are equally severe: Toward severity grading of individual criteria in college drinkers. *Psychology of addictive behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *33*(1), 35–49. https://doi.org/10.1037/adb0000443

Bonifay, W., & Cai, L. (2017). On the Complexity of Item Response Theory Models. *Multivariate behavioral research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Buzbas, E. O., Devezer, B., & Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science,* 10: 221042. http://doi.org/10.1098/rsos.221042

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Cudeck, R. & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109*(3), 512-519.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science, 8*(3), 200805. https://doi.org/10.1098/rsos.200805

Dyson, F. J. (1999). *The sun, the genome, and the Internet: Tools of scientific revolutions.* New York, NY: Oxford University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Evans. M., & Jang, G. H. (2010). Invariant P-values for model checking. *Annals of Statistics, 38*(1), 512-525. https://doi.org/10.1214/09-AOS727

Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis, 1*(4), 893-914. https://doi.org/10.1214/06-ba129

Falk, C. F., & Muthukrishna, M. (2023). Parsimony in model selection: Tools for assessing fit propensity. *Psychological methods*, *28*(1), 123–136. https://doi.org/10.1037/met0000422

Fernández de la Cruz L, Vidal-Ribas P, Zahreddine N, Mathiassen B, Brøndbo PH, Simonoff E, Goodman R, & Stringaris A. (2018). Should clinicians split or lump psychiatric symptoms? The structure of psychopathology in two large pediatric clinical samples from England and Norway. *Child Psychiatry Hum Dev*, *49*(4), 607-620. https://doi.org/10.1007/s10578-017-0777-1

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271-288.

Fryar, C. D., Carroll, M. D., Gu, Q., Afful, J., & Ogden, C. L. (2021). Anthropometric Reference Data for Children and Adults: United States, 2015-2018. *Vital & health statistics. Series 3, Analytical and epidemiological studies*, (36), 1–44.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, *6*, 721-741.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733-760.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis (3rd Ed.)*. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Giuntoli, L., Condini, F., Ceccarini, F., Huta, V., & Vidotto, G. (2021). The different roles of hedonic and eudaimonic motives for activities in predicting functioning and well-being experiences. *Journal of Happiness Studies*, *22*, 1657-1671.

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology, 40*(7), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 83-100.

Haig, B. D. (2022). Understanding replication in a way that is true to science. *Review of General Psychology, 26*(2), 224–240. https://doi.org/10.1177/10892680211046514

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res., 15*(1), 1593-1623.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

E. T. Jaynes (2003). *Probability Theory: The Logic of Science.* Cambridge, England: Cambridge University Press.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry*, *51*(1), 8–19. https://doi.org/10.1001/archpsyc.1994.03950010008002

Kessler, R.C., & Merikangas, K.R. (2004), The National Comorbidity Survey Replication (NCS-R): Background and aims. *International Journal of Methods in Psychiatric Research, 13*, 60-68. https://doi.org/10.1002/mpr.166

Kessler, R.C., & Üstün, T.B. (2004), The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research, 13*, 93-121. https://doi.org/10.1002/mpr.168

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology?. *Theory & Psychology*, *24*(3), 326-338.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Krueger R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, *56*(10), 921–926. https://doi.org/10.1001/archpsyc.56.10.921

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4), 722-752.

Lane, S. P., Steinley, D., & Sher, K. J. (2016). Meta-analysis of DSM alcohol use disorder criteria severities: structural consistency is only 'skin deep.' *Psychological Medicine, 46*, 1769–1784.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3–4), 141–153. https://doi.org/10.1007/s42113-019-00029-y

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Mayo, D. G. (1997). Severe tests, arguing from error, and methodological underdetermination. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *86*(3), 243-266.

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician, 73*, 99-105.

Merkle, E.C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, 85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, *2*(1), 28–34. https://doi.org/10.1007/s42113-018-0019-z

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 116–162). New York: Chapman & Hall/CRC Press.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and 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 BA, Errington TM. (2020). What is replication? *PLoS Biol*, *18*(3), e3000691. https://doi.org10.1371/journal.pbio.3000691

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 1–8.

Osborne, J. W. & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research, and Evaluation, 17*(15), 1-8. https://doi.org/10.7275/h0bd-4d11

Paruzel-Czachura, M., & Blukacz, M. (2021). How relevant for you is to be a moral person? Polish validation of the Self-Importance of Moral Identity Scale. *Plos one*, *16*(8), e0255386.

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421-425.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41*(3), 227-259.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295x.107.2.358

Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review, 109*(3), 605–607. doi:10.1037//0033-295X.109.3.605.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rosseel Y (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics, 9*(1), 130-134.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151-1172.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of abnormal psychology*, *112*(4), 578–598. https://doi.org/10.1037/0021-843X.112.4.578

Vanpaemel, W. (2019). The really risky registered modeling report: Incentivizing strong tests and HONEST modeling in cognitive science. *Computational Brain & Behavior*, *2*(3–4), 218–222. https://doi.org/10.1007/s42113-019-00056-9

Vanpaemel, W. (2019). The really risky registered modeling report: Incentivizing strong tests and HONEST modeling in cognitive science. *Computational Brain & Behavior*, *2*(3-4), 218-222.

Vanpaemel W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological review*, *127*(1), 136–145. https://doi.org/10.1037/rev0000167

van Zundert, C., Somer, E., & Miocevic, M. (2022). Prior predictive checks for the method of covariances in Bayesian mediation analysis. Structural Equation Modeling, 29, 428–437. https://doi.org/10.1080/10705511.2021.1977648

Villarreal, M., Etz, A., & Lee, M. D. (2023). Evaluating the complexity and falsifiability of psychological models. *Psychological review*, 10.1037/rev0000421. Advance online publication. https://doi.org/10.1037/rev0000421

Waller, N. G., & Meehl, P. E. (2002). Risky tests, verisimilitude, and path analysis. *Psychological Methods*, *7*(3), 323–337. https://doi.org/10.1037/1082-989X.7.3.323

Whiteman, S. E., Kramer, L. B., Silverstein, M. W., Witte, T. K., & Weathers, F. W. (2022). Evaluating the factor structure of the Posttraumatic Cognitions Inventory. *Assessment, 29*(2), 128-135.

Winter, S. D., & Depaoli, S. (2023). Illustrating the Value of Prior Predictive Checking for Bayesian Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-22.

Wittchen H. U. (1994). Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): A critical review. *Journal of Psychiatric Research*, *28*(1), 57–84. https://doi.org/10.1016/0022-3956(94)90036-1

Youyou, W., Yang, Y., & Uzzi, B. (2023). A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(6), e2208863120. https://doi.org/10.1073/pnas.2208863120

Zondervan-Zwijnenburg, M. (2019, March 4). How to Test Replication for Structural Equation Models. https://doi.org/10.31234/osf.io/uvh5s

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development ISSN*, *14*, 305–320. https://doi.org/10.1080/15427609.2017.1370966

Appendix

PrPMC technical details. More technically, the prior predictive distribution is generated by evaluating the marginal likelihood *without* including the observed data. In other words, this simulated distribution comes from a Bayesian analysis for which the iterative output for parameter θ is only affected by the prior distribution $p(\theta)$: $\theta^{\text{pred}} \sim p(\theta)$. The resulting distribution is also known as the *prior pushforward distribution*, which can be used to simulate data $y^{pred} \sim p\left(y \mid \theta^{\text{pred}}\right)$ according to the sampling distribution, given the simulated model parameters (i.e., prior pushforward distributions). The result is a simulation from the joint distribution $\left(y^{\text{pred}}, \theta^{\text{pred}}\right) \sim p(y, \theta)$, which means that $y^{\text{pred}} \sim p(y)$ is a simulation from the prior predictive distribution.

In practice, the prior pushforward samples θ^{pred} are generated through iterative algorithms such as the Markov Chain Monte Carlo algorithm with the Gibbs sampler (Geman & Geman, 1984) or the Hamiltonian Monte Carlo (Neal, 2011) algorithm with the No-U-Turn Sampler (Hoffman & Gelman, 2014). For each sample of θ^{pred}, a corresponding prior predictive sample y^{pred} can be generated, resulting in a distribution of all possible samples that could occur given the model and prior specification (Van de Schoot et al., 2021). The prior pushforward distribution and prior predictive samples can be used to assess the extent to which the prior and likelihood generate data that are in line with our expectations (e.g., what replication data *should* look like, given the original findings). Ultimately, we want to ensure that most of the prior distribution covers areas of the parameter space that correspond to reasonable data generating processes, given our expectations (Gelman et al., 2017).

We used the R package *blavaan* (Merkle & Rosseel, 2018) to generate the prior pushforward and prior predictive samples for each sample size and prior specification

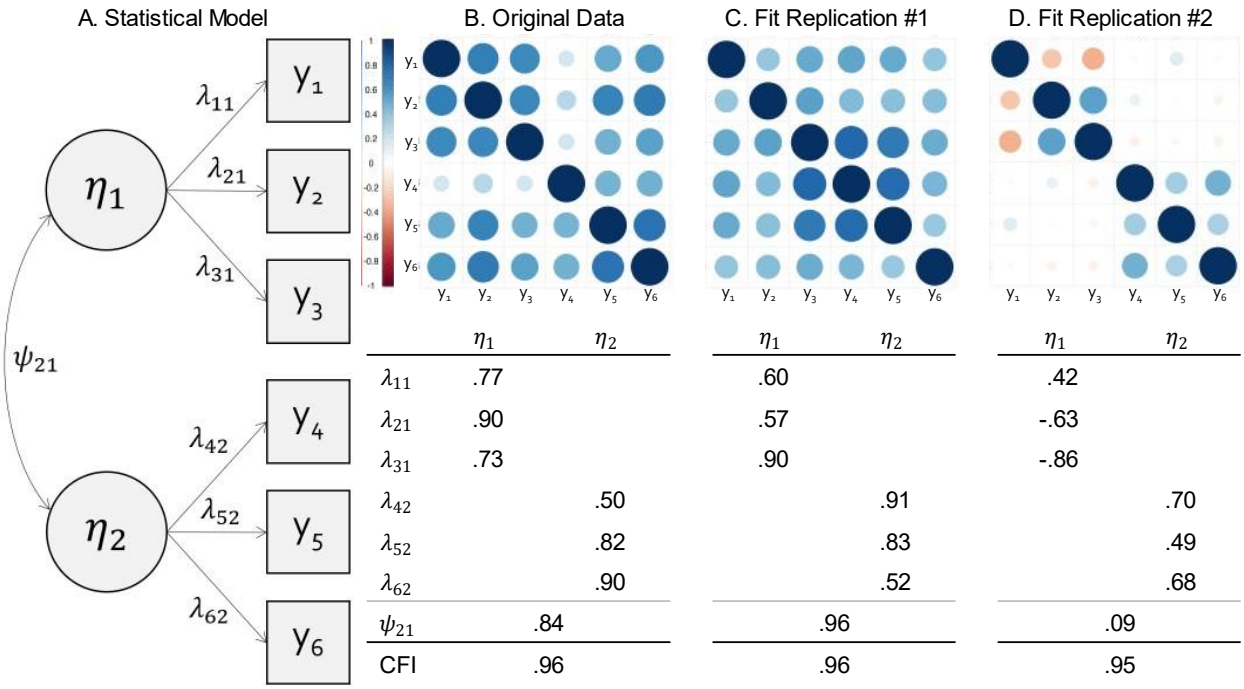
combination. For each analysis, we requested two chains consisting of 1000 burn-in samples and 1000 prior pushforward samples. This number of samples was sufficient to result in stable prior pushforward chains (i.e., all R-hat statistics < 1.01). From the 2000 pushforward samples, we generated 2000 prior predictive samples for use in assessing the similarity of the replication data to the original (model-implied) data.

Similarity of Response Proportions. To demonstrate the capabilities of PrPMC to evaluate any feature of the data, we inspected a second set of test statistics: the proportions of respondents who were scored as exhibiting each mental disorder. In Figure A1, the horizontal lines represent $T(y)$, which are the response proportions in the replication data, and the bars indicate $T(y^{\text{pred}})$, which are the prior predictive distributions based on said proportions in the original data. This analysis shows that the original and replication data may be more dissimilar with respect to one data feature, but more similar with respect to another: Where the pairwise correlations among indicators were less successful at approximate empirical replication, the response proportions always hit the approximate replication targets (and, of course, the much easier fit replication targets). For close empirical replication, only drug dependence passed this demanding test.

Table 1. *Original study estimates and corresponding distributions for close and approximate empirical replication and fit replication.*

	Original Study			Replication Target Prior Distributions		
	Estimate	Std. Err.	β	Close	Approximate	Fit
Loadings						
Externalizing						
Alcohol dependence	1.00	–	.79	N(.79, .05)	N(.79, .50)	N(0, 5)
Drug dependence	1.18	.16	.84	N(.84, .05)	N(.84, .50)	N(0, 5)
Conduct disorder	.60	.06	.61	N(.61, .05)	N(.61, .50)	N(0, 5)
Distress						
Major depression	1.00	–	.70	N(.70, .05)	N(.70, .50)	N(0, 5)
Dysthymia	1.28	.14	.78	N(.78, .05)	N(.78, .50)	N(0, 5)
Generalized anxiety	1.28	.13	.79	N(.79, .05)	N(.79, .50)	N(0, 5)
Fear						
Agoraphobia	1.00	–	.60	N(.60, .05)	N(.60, .50)	N(0, 5)
Social anxiety	1.43	.14	.73	N(.73, .05)	N(.73, .50)	N(0, 5)
Specific phobia	1.48	.14	.74	N(.74, .05)	N(.74, .50)	N(0, 5)
Panic disorder	1.32	.16	.70	N(.70, .05)	N(.70, .50)	N(0, 5)
Thresholds						
Alcohol dependence	1.69	.08	−	N(1.69, .05)	N(1.69, .50)	N(0, 3)
Drug dependence	2.58	.16	−	N(2.58, .05)	N(2.58, .50)	N(0, 3)
Conduct disorder	1.46	.04	−	N(1.46, .05)	N(1.46, .50)	N(0, 3)
Major depression	1.40	.05	−	N(1.40, .05)	N(1.40, .50)	N(0, 3)
Dysthymia	2.37	.10	−	N(2.37, .05)	N(2.37, .50)	N(0, 3)
Generalized anxiety	2.64	.12	−	N(2.64, .05)	N(2.64, .50)	N(0, 3)
Agoraphobia	2.09	.06	−	N(2.09, .05)	N(2.09, .50)	N(0, 3)
Social anxiety	1.64	.06	−	N(1.64, .05)	N(1.64, .50)	N(0, 3)
Specific phobia	1.84	.07	−	N(1.84, .05)	N(1.84, .50)	N(0, 3)
Panic disorder	2.57	.11	−	N(2.57, .05)	N(2.57, .50)	N(0, 3)
Correlations						
Externalizing–Distress	.50	.05	.40	Beta(97.7, 42.8)	Beta(16.5, 7.7)	LKJ(1)
Externalizing–Fear	.37	.05	.39	Beta(98.5, 43.9)	Beta(16.0, 7.6)	LKJ(1)
Distress–Fear	.46	.05	.63	Beta(96.3, 22.9)	Beta(24.5, 6.4)	LKJ(1)

Note. Original study N = 8,098. β = standardized estimate; N(·) = Normal distribution; LKJ(·) = Lewandowski-Kurowicka-Joe distribution.



	B. Original Data		C. Fit Replication #1		D. Fit Replication #2	
	η_1	η_2	η_1	η_2	η_1	η_2
λ_{11}	.77		.60		.42	
λ_{21}	.90		.57		-.63	
λ_{31}	.73		.90		-.86	
λ_{42}		.50		.91		.70
λ_{52}		.82		.83		.49
λ_{62}		.90		.52		.68
ψ_{21}	.84		.96		.09	
CFI	.96		.96		.95	

Figure 1. The confirmatory factor model in Panel A will fit well (comparative fit index (CFI) ≥ .95) to all three correlation matrices, despite considerable differences in the data patterns and parameter estimates. Data matrices are available on OSF (https://osf.io/q6rvf/).

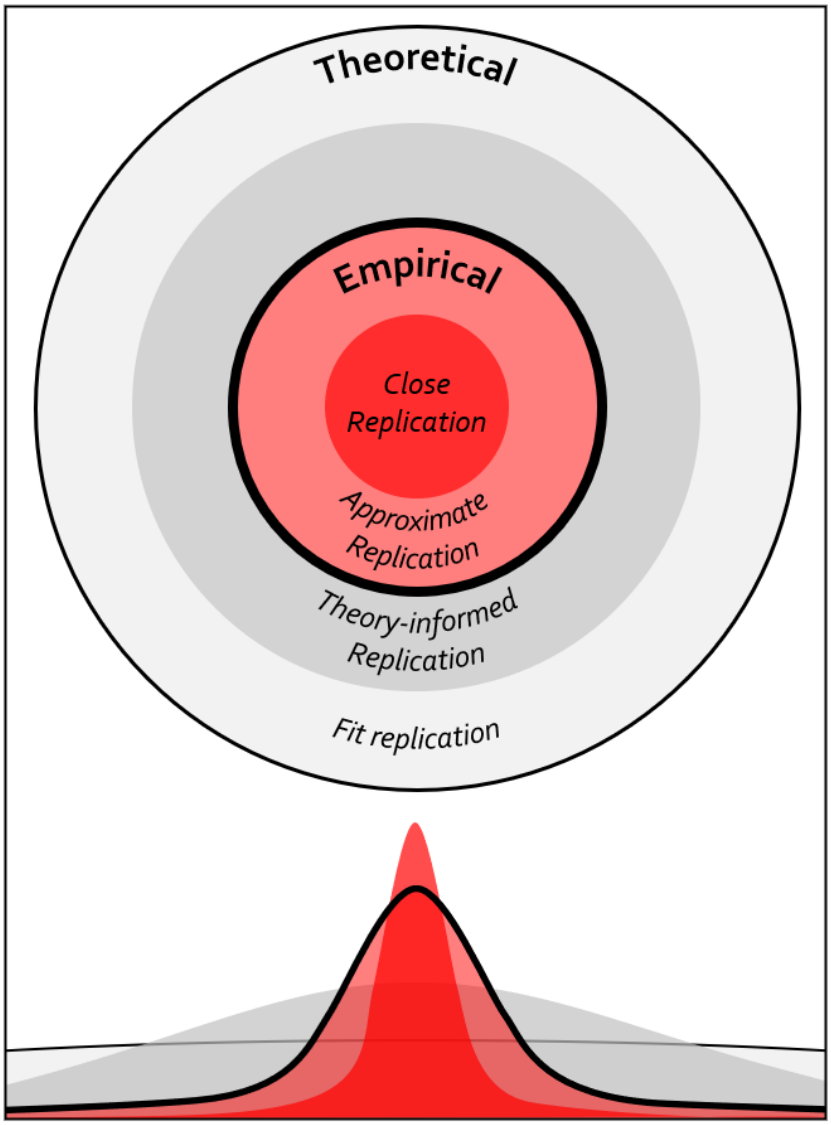


Figure 2. Target-setting is essential when investigating statistical model replication. To improve upon simple replication of fit, researchers must aim at the more meaningful inner regions of the target. Doing so will entail riskier testing of the data and/or model parameters, which can be accomplished by using Bayesian prior predictive similarity checking with increasingly informative prior distributions, as depicted at the bottom of this figure. Theory-informed replication requires, at a minimum, checking specific hypotheses about the data and/or model parameters. Approximate or close empirical replication requires checking that the replicated data and/or model parameters approximately or closely resemble those of the original study.

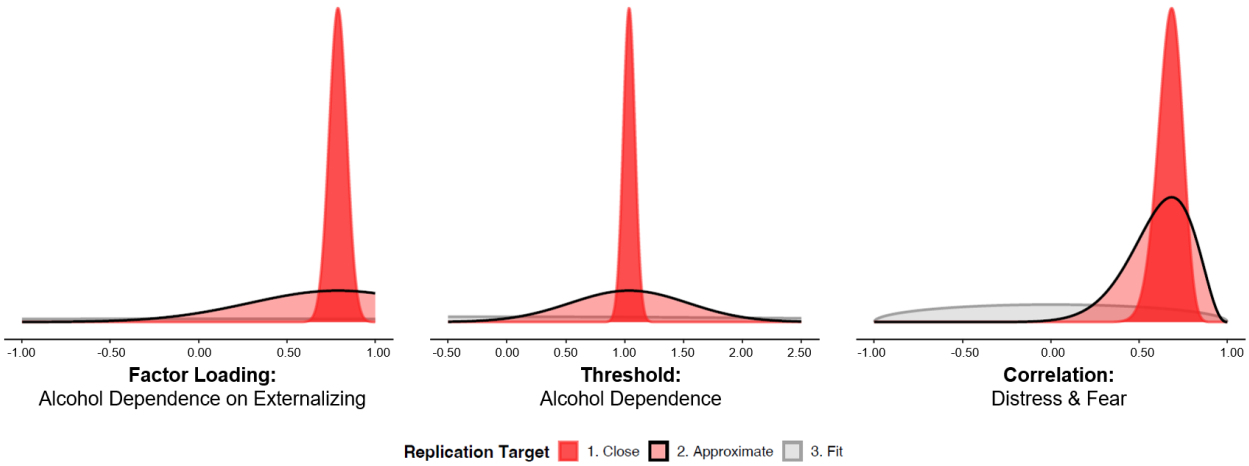


Figure 3. Example prior specifications when targeting close (deep red) or approximate (light red) empirical replication, in contrast to targeting model fit (gray) without checking the similarity between the data and parameter estimates of the original and replication studies. For ease of visualization, *x*-axes are truncated to a reasonable range for each prior. See Table 1 for detailed prior specifications.

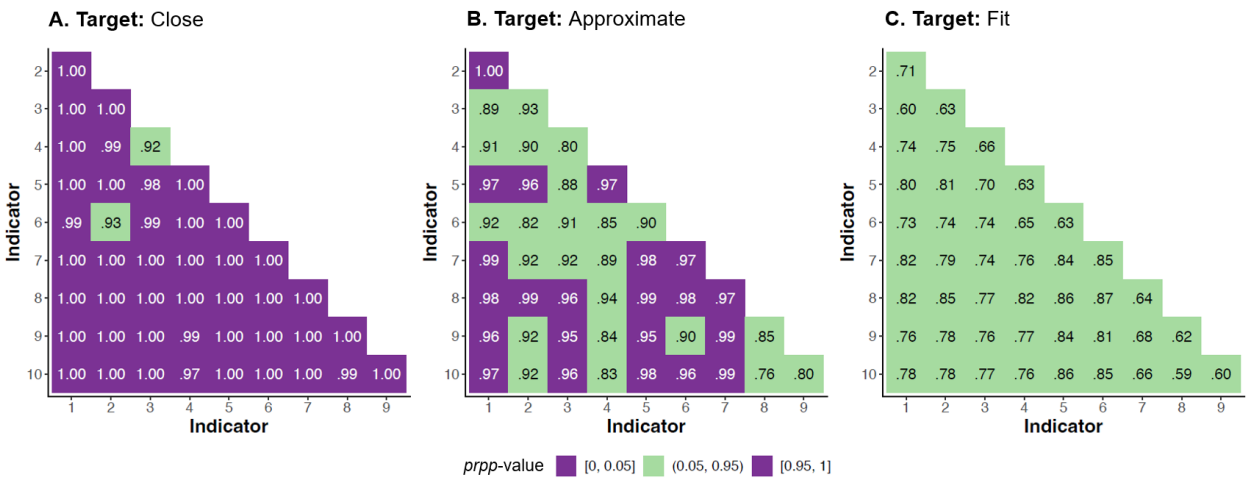


Figure 4. Prior predictive checks of the similarity between original and replication data correlation matrices when targeting close (left) or approximate (center) empirical replication versus fit replication (right). Green cells indicate correlations within the replication data that align with the bulk of the corresponding prior predictive distribution (i.e., *prpp*-values between .05 and .95). Purple cells indicate replication data correlations at the extremes of the corresponding prior predictive distribution (i.e., *prpp* $\leq$ .05 or $\geq$ .95). Indicators: 1 = alcohol dependence; 2 = drug dependence; 3 = conduct disorder; 4 = major depression; 5 = dysthymia; 6 = generalized anxiety disorder; 7 = agoraphobia; 8 = social anxiety disorder; 9 = specific phobia; 10 = panic disorder.

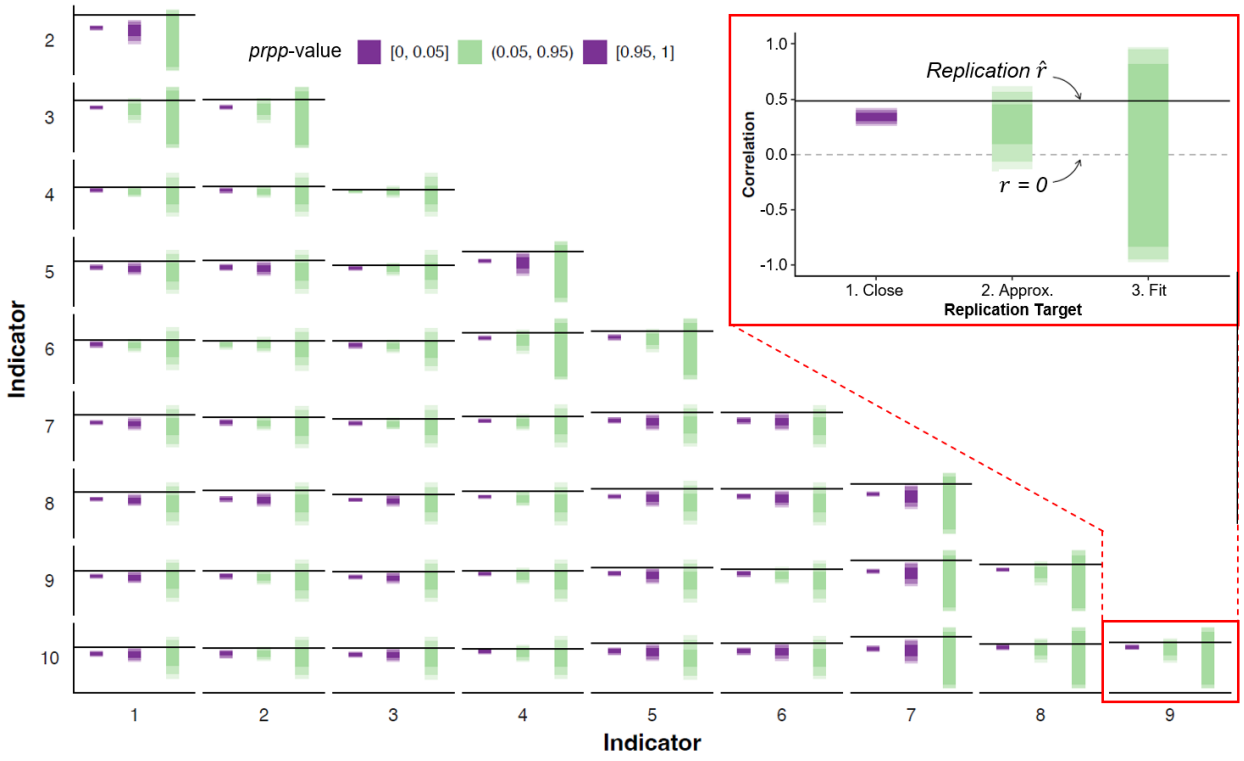


Figure 5. Prior predictive similarity checks of the replication data correlations. Within each plot, the horizontal line denotes the correlation between each pair of mental disorders in the replication data, and the bars depict targets of close (left) and approximate (center) empirical replication and fit replication (right). Green indicates that the replication hit the target (*prpp*-values between .05 and .95) and purple indicates that the replication missed the target (*prpp* $\leq$ .05 or $\geq$ .95). Indicators: 1 = alcohol dependence; 2 = drug dependence; 3 = conduct disorder; 4 = major depression; 5 = dysthymia; 6 = generalized anxiety disorder; 7 = agoraphobia; 8 = social anxiety disorder; 9 = specific phobia; 10 = panic disorder.

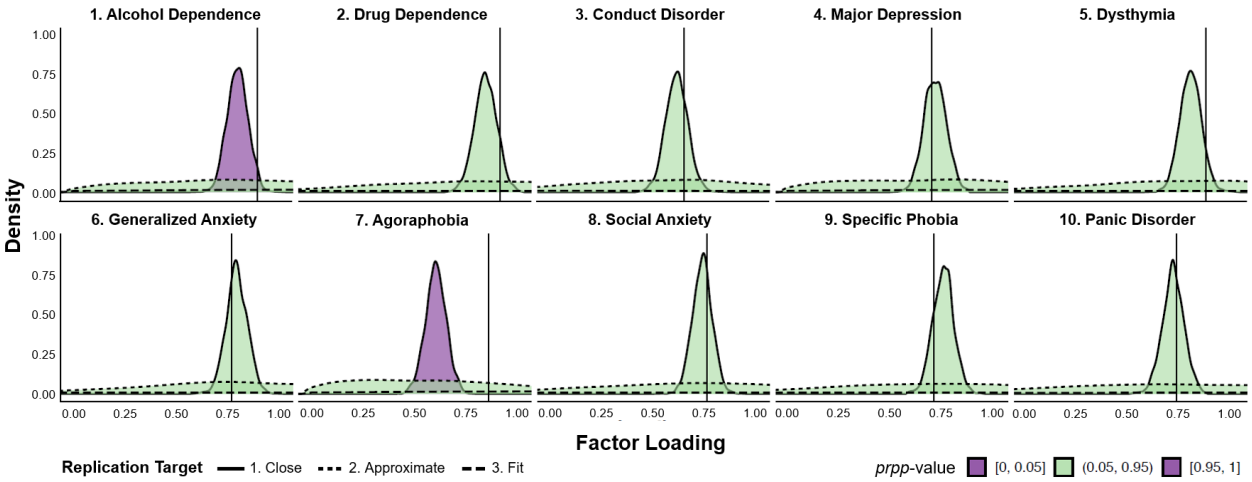


Figure 6. Prior predictive similarity checks of the replication factor loadings when targeting close (solid) and approximate (dotted) empirical replication and fit replication (dashed). The vertical lines denote the parameter estimates from the replication data. Green distributions indicate replication loadings that hit the target (*prpp* $>$ .05 and $<$ .95) and purple distributions indicate replication loadings that missed the target (*prpp* $\leq$ .05 or $\geq$ .95).

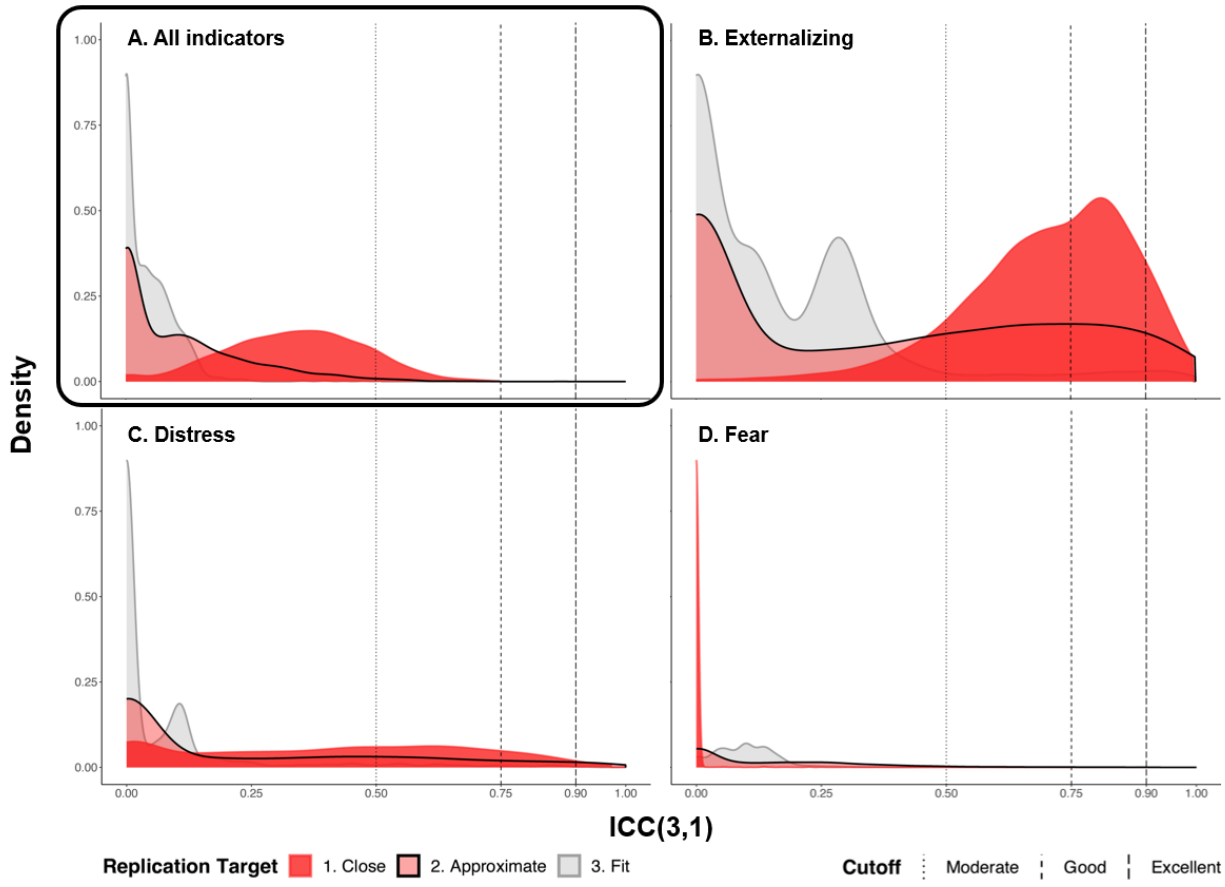


Figure 7. Prior predictive similarity checks of the loading patterns across all indicators (Panel A) and within each factor (Panels B-D), when targeting close (deep red) or approximate (light red) empirical replication and fit replication (gray). Pattern consistency was indexed by the ICC(3,1) intraclass correlation coefficient (Shrout & Fleiss, 1979). The vertical lines correspond to the recommendations of Koo and Li (2016), who proposed ICC cutoff values of .50, .75, and .90 as indicative of moderate, good, and excellent consistency, respectively. Externalizing = indicators 1-3 (alcohol dependence, drug dependence, conduct disorder); Distress = indicators 4-6 (major depression, dysthymia, generalized anxiety); Fear = indicators 7-10 (agoraphobia, social anxiety, specific phobia, panic disorder).

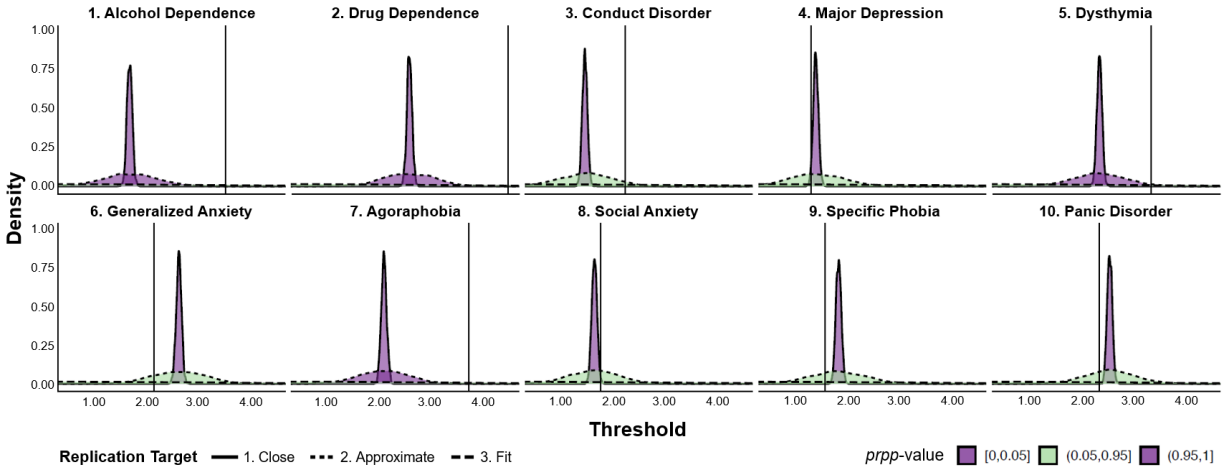


Figure 8. Prior predictive similarity checks of the replication thresholds when targeting close (solid) and approximate (dotted) empirical replication and fit replication (dashed). The vertical lines denote the parameter estimates from the replication data. Green distributions indicate replication loadings that hit the target ($prpp > .05$ and $< .95$) and purple distributions indicate replication loadings that missed the target ($prpp \leq .05$ or $\geq .95$).

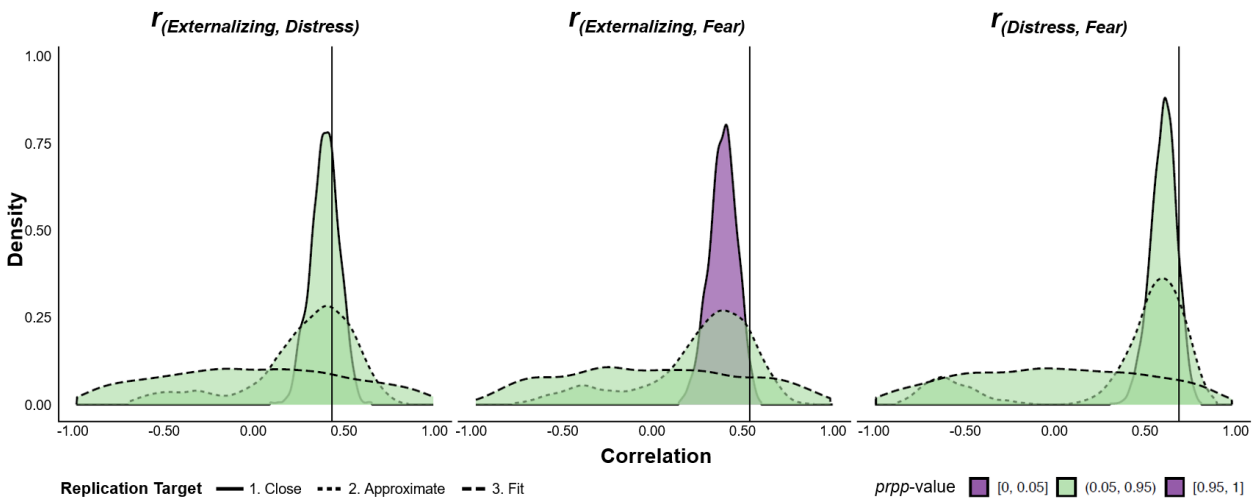


Figure 9. Prior predictive similarity checks of the factor correlations when targeting close (solid) and approximate (dotted) empirical replication and fit replication (dashed). The vertical lines denote the factor correlations estimated from the replication data: Green distributions indicate replication factor correlations that hit the target (*prpp* > .05 and < .95) and purple distributions indicate replication factor correlations that missed the target (*prpp* ≤ .05 or ≥ .95). Externalizing = indicators 1-3 (alcohol dependence, drug dependence, conduct disorder); Distress = indicators 4-6 (major depression, dysthymia, generalized anxiety); Fear = indicators 7-10 (agoraphobia, social anxiety, specific phobia, panic disorder).

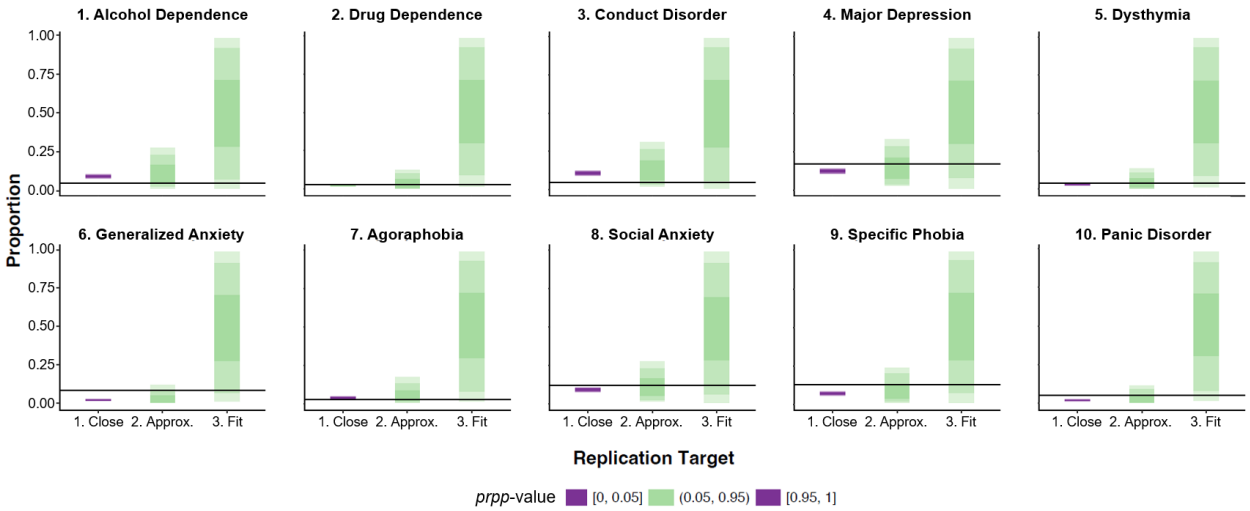


Figure A1. Prior predictive similarity checks of the indicator response proportions in the original and replication data. Within each plot, the horizontal line denotes prop($\hat{y}_i$ = *present*) of each mental disorder in the replication data, and the bars depict targets of close (left) and approximate (center) empirical replication and fit replication (right). Green indicates that the replication hit the target and purple indicates that the replication missed the target.