Testing Small Variance Priors Using Prior-Posterior Predictive P-values

Herbert Hoijtink

Department of Methodology and Statistics, Utrecht University and CITO Institute for Educational Measurement

Rens van de Schoot

Department of Methodology and Statistics, Utrecht University and North-West University, Optentia Research Program, South Africa

Author Note

Correspondence concerning this article should be addressed to Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands. E-mail: H.Hoijtink@uu.nl. The first author is supported by the Consortium on Individual Development (CID) which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). The second author is supported by a VIDI grant from the Netherlands Organization for Scientific Research (grant number NWO-VIDI-452-14-006). We are indebted to one of the anonymous reviewers for many good ideas and suggestions. There has been no prior dissemination of the ideas expressed in this paper before it was submitted for publication to Psychological Methods.

Abstract

Muthen and Asparouhov (2012) propose to evaluate model fit in structural equation models based on approximate (using small variance priors) instead of exact equality of (combinations of) parameters to zero. This is an important development that adequately addresses Cohen's (1994) "The earth is round ($p<.05$)", which stresses that point null-hypotheses are so precise that small and irrelevant differences from the null-hypothesis may lead to their rejection. It is tempting to evaluate small variance priors using readily available approaches like the posterior predictive p-value and the DIC. However, as will be shown, both are not suited for the evaluation of models based on small variance priors. In this paper a well behaving alternative, the prior-posterior predictive p-value, will be introduced. It will be shown that it is consistent, the distributions under the null and alternative hypotheses will be elaborated, and it will be applied to testing whether the difference between two means and the size of a correlation are relevantly different from zero.

Keywords: DIC, Posterior Predictive P-value, Prior-Posterior Predictive P-value, SEM, Small Variance Prior

Testing Small Variance Priors Using Prior-Posterior Predictive P-values

Introduction

The title of Cohen's 1994 paper "The Earth is Round: p < .05 " stresses that small and irrelevant deviations from a null-hypothesis may result in a rejection of the null-hypothesis. Muthen and Asparouhov (2012) nicely address this problem by replacing the "exactly equal to zero" constraints in structural equation models by approximate equality constraints using, so-called, small variance priors. Their idea is very useful in the context of, for example, confirmatory factor analysis, where such small variance priors enable the replacement of cross-loadings that are fixed at zero (see Muthen and Asparouhov, 2012) by cross loadings that are allowed wiggle room around zero. In the context of the evaluation of measurement invariance they allow replacing strict by approximate measurement invariance (see, e.g., Van de Schoot et al. 2013). The interested reader is also refferred to Maccallum, Edwards, and Cai (2012) and Rindskopf (2012) for further discussions of small variance priors.

The following simple regression model (which will be used to illustrate matters throughout this paper) illustrates replacing an exactly equal to zero constraint by a small variance prior. Let

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, ..., N$. The traditional null-hypothesis (a.k.a. the Earth is exactly round) is

$$H_0 : \beta = 0. \tag{2}$$

Applying the proposal by Muthen and Asparouhov (2012) to the simple regression model renders

$$H_\approx : \beta \sim \mathcal{N}(0, \tau^2), \tag{3}$$

that is, Muthen and Asparouhov (2012) place a normal prior distribution on β that is centered around zero and has a small variance τ^2 such that small deviations from zero are allowed (a.k.a. the Earth is approximately round). Like Muthen and Asparouhov (2012) in some of their examples, in this paper $\tau^2 = .01$ will be used. This implies that 95% of the prior probability mass for β is between -.2 and +.2. In the simple regression model small variance priors enable testing the null-hypothesis "the deviation of β from zero is irrelevant" versus the alternative "the deviation of β from zero is relevant". Note that we will use standard non-informative prior distributions for the other parameters in the model: $\alpha \sim \mathcal{N}(0, 1000000)$ and $\sigma^2 \sim \Gamma^{-1}(-1, 0)$. The three prior distributions will be denoted by $h(\beta)$, $h(\alpha)$, and $h(\sigma^2)$, respectively.

It is as to yet an unresolved issue how to test hypotheses based on small variance priors (like $H_{\approx}$ from Equation 3). One approach is to compute the posterior predictive p-value (Meng, 1994) which will be elaborated in the next section. Another approach is based on model comparison by means of the deviance information criterion (DIC, Spiegelhalter, Best, Carlin and Van Der Linde, 2002) which will also be elaborated in the next section. Bayesian software provides the posterior predictive p-value and/or the DIC by default: Mplus (Muthen and Muthen, 1998-2015), Blavaan (Merkle and Rosseel, 2015), and AMOS (http://www.spss.com.hk/amos/), render the posterior predictive p-value and DIC; WinBugs/OpenBugs (Lunn, Thomas, Best, and Spiegelhalter, 2000), and JAGS (http://mcmc-jags.sourceforge.net/), render the DIC. It is therefore tempting to also use these tools to test hypotheses based on small variance priors. Indeed, in a large systematic review on the use of Bayes statistics in Psychological research by Van de Schoot et al. (2016) three empirical papers were identified that tested small variance priors for cross-loadings (Falkenstrom, Hatcher, and Holmqvist, 2015; Golay, et al. 2013; Ryoo et al. 2015) and five papers that used small variance priors to test measurement invariance (Bujacz et al., 2014; Chiorri et al., 2014; Cieciuch et al., 2014; Jackson et al., 2014; Zercher et al., 2015). The posterior predictive p-value and the DIC were also used in simulation studies (Kelcey et al., 2014; Muthen and Asparouhov, 2012; Strohmeyer et al., 2015; Van de Schoot et al., 2013). The interested reader is also referred to Appendix A of Asparohov, Muthen and Morin (2015). They give a step by step description of the evaluation of models based on small variance priors in which there is no role for the posterior predictive p-value or the DIC.

In the current paper it will be shown that the posterior predictive p-value and the DIC can not be used to evaluate small variance priors. An alternative will be proposed, the prior-posterior predictive p-value, which will shown to be well suited for the evaluation of models based on small variance priors. To illustrate the approach proposed, small variance priors for a correlation and the difference between two means will be tested. This paper will be concluded with a discussion of the main results and their importance for the evaluation of small variance priors in the context of full fledged structural equation models.

Evaluating Small Variance Priors Using the Posterior Predictive P-value and the DIC

The Posterior Predictive P-value and the DIC

In this paper the posterior predictive p-value Meng (1994) presented in Muthen and Asparouhov (2012) and implemented in Mplus (Muthen and Muthen, 1998-2015) will be used. It is based on a discrepancy measure (the likelihood ratio statistic) comparing $H_{\approx}$ with H_u (the saturated model). Let $\theta = [\alpha, \beta, \sigma^2]$, that is, θ contains the parameters from the regression model presented in Equation 1, then:

$$p_{\approx} = P(D_{rep} > D|\boldsymbol{y}, \boldsymbol{x}, H_{\approx}) = \int_{\theta} P(D_{rep} > D|\theta) g(\theta|\boldsymbol{y}, \boldsymbol{x}, H_{\approx}) d\theta, \tag{4}$$

where, $g(.)$ denotes the posterior distribution of θ based on the small variance prior from Equation 3:

$$g(.) \propto f(\boldsymbol{y} \mid \boldsymbol{x}, \theta) h(\beta) h(\alpha) h(\sigma^2) = \frac{-N-1}{2} (\log|\Sigma(\theta)| + tr[S\Sigma^{-1}(\theta)]) h(\beta) h(\alpha) h(\sigma^2). \tag{5}$$

Note that, the discrepancy measure

$$D_{rep} = \log|\Sigma(\theta)| - \log|S_{rep}| + \text{tr}[S_{rep}\Sigma^{-1}(\theta)] - k, \tag{6}$$

and

$$D = \log|\Sigma(\theta)| - \log|S| + \text{tr}[S\Sigma^{-1}(\theta)] - k, \tag{7}$$

with $k = 2$ denoting the number of observed variables. The unbiased sample covariance matrix S contains the variances and covariance of $\boldsymbol{y}$ and $\boldsymbol{x}$

$$S = \begin{bmatrix} s^2_{\boldsymbol{y}} & s_{\boldsymbol{yx}} \\ s_{\boldsymbol{yx}} & s^2_{\boldsymbol{x}} \end{bmatrix}, \tag{8}$$

the unbiased sample covariance matrix S_{rep} contains the variances and covariance of $\boldsymbol{y}_{rep}$ and $\boldsymbol{x}$

$$S_{rep} = \begin{bmatrix} s^2_{\boldsymbol{y}_{rep}} & s_{\boldsymbol{y}_{rep}\boldsymbol{x}} \\ s_{\boldsymbol{y}_{rep}\boldsymbol{x}} & s^2_{\boldsymbol{x}} \end{bmatrix}, \tag{9}$$

where $\boldsymbol{y}_{rep}$ is generated using the simple regression model, θ, and $\boldsymbol{x}$, and the covariance matrix implied by θ

$$\Sigma(\theta) = \begin{bmatrix} \beta^2 s^2_{\boldsymbol{x}} + \sigma^2 & \beta s^2_{\boldsymbol{x}} \\ \beta s^2_{\boldsymbol{x}} & s^2_{\boldsymbol{x}} \end{bmatrix}. \tag{10}$$

Besides the posterior predictive p-value we will use the DIC (Spiegelhalter, Best, Carlin and Van Der Linde, 2002) as presented by Asparouhov, Muthen, and Morin (2015) and implemented in Mplus (Muthen and Muthen, 1998-2015):

$$DIC_{\approx} = \bar{C}_{\approx} + q_{\approx}, \tag{11}$$

where,

$$\bar{C}_{\approx} = \int_{\theta} -2\log f(\boldsymbol{y} \mid \boldsymbol{x}, \theta) g(\theta \mid \boldsymbol{y}, \boldsymbol{x}, H_{\approx}) d\theta, \tag{12}$$

the estimated number of parameters

$$q_{\approx} = \bar{C}_{\approx} - 2\log f(\boldsymbol{y} \mid \boldsymbol{x}, \hat{\theta}), \tag{13}$$

and, $\hat{\theta}$ denotes the mean of the posterior distribution $g(.)$.

Note that, the notation $p_{\approx}$ and $DIC_{\approx}$ is used to stress that the posterior predictive p-value and DIC are computed using a posterior distribution $g(.)$ that is based on small variance prior. It is furthermore important to note that, the influence of the small variance prior can only enter the computation of the posterior predictive p-value and the DIC through this posterior distribution. This has consequences that will be illustrated in the next two subsections.

Performance of $p_{\approx}$ and $DIC_{\approx}$ for the Evaluation of the Small Variance Prior Used in the Simple Regression Model

To illustrate that $p_{\approx}$ from Equation 4 and $DIC_{\approx}$ from Equation 11 provide poor evaluations of the fit of $H_{\approx}$, four data sets where generated with sample sizes N of 25, 100, 1000, and 10000, respectively, such that the maximum likelihood estimates $\hat{\alpha} = 0$, $\hat{\beta} = .707$, $\hat{\sigma}^2 = .50$, and the sample mean and variance of $\boldsymbol{x}$ are 0 and 1, respectively (data generated using BIEMS, Mulder, Hoijtink, and de Leeuw, 2012). Clearly, the estimate of $\hat{\beta} = .707$ is *not* in agreement with $H_{\approx}$ and a substantial amount of the variance of $\boldsymbol{y}$ (50%) is explained by $\boldsymbol{x}$. Consequently, with increasing sample sizes the evidence against $H_{\approx}$ should become stronger, that is, $p_{\approx}$ should become smaller. However, as can be seen in the third column of Table 1 (computed using Mplus version 7.4, Muthen and Muthen, 1998-2015, with fbiterations=100000) the posterior predictive p-value does not decrease but actually the opposite happens! This is caused by the fact that with increasing sample sizes the data dominate the prior, stated otherwise, with increasing sample sizes the influence of the prior on the posterior $g(.)$ disappears.

In the second column of Table 1 the $p_{\approx}$'s are displayed that result if the prior distribution $\beta \sim \mathcal{N}(0, .01)$ is replaced by the uninformative prior $\beta \sim \mathcal{N}(0, 1000000)$, that is, irrespective of the value of N, there is no prior influence on the posterior distribution. What is tested here is the fit to the data of the simple regression model. Since the data were generated using the simple regression model, it is not a surprise that the fit is good (i.e., the $p_{\approx}$'s are close to .50). As can be seen comparing the second and third columns, with increasing N both posterior predictive p-values converge towards each other. This implies that with increasing N both posterior predictive p-values test the fit of the simple regression model and nothing else.

In the last four columns of Table 1, $DIC_{\approx}$ and $q_{\approx}$ are presented for $\tau^2 = 1000000$ and $\tau^2 = .01$, respectively. Since the data are *not* in agreement with $\tau^2 = .01$, with increasing sample size the difference between both DIC's should increase (the data generated are not conflicting with $\beta \sim \mathcal{N}(0, 1000000)$ but are conflicting with $\beta \sim \mathcal{N}(0, .01)$). However, as can be seen, the opposite happens. This too is caused by the fact that with increasing N the data increasingly dominate the prior and therefore that the influence of the prior on the posterior distribution and therefore on the outcome of the model evaluation vanishes. This is further illustrated comparing the estimated number of parameters $q_{\approx}$. For smaller sample sizes the small variance prior imposes restrictions on the parameter space and therefore the estimated number of parameters for $\tau^2 = .01$ is smaller than for $\tau^2 = 1000000$. However, with increasing sample size the influence of the prior vanishes and the estimated number of parameters converges to 3, which is number of parameters used to specify the regression model in Equation 1.

These observations do not depend on the fact that $\boldsymbol{x}$ is not modeled in Equation 1. If the model is extended with with $x_i \sim \mathcal{N}(\mu, \sigma^2_{\mu})$ for $i = 1, ..., N$, $h(\mu) \sim \mathcal{N}(0, 1000000)$, and $h(\sigma^2_{\mu}) \sim \Gamma^{-1}(-1, 0)$, Mplus renders the results displayed in Table 2. As can be seen, the conclusions are identical to the conclusions obtained from Table 1.

Finally, Table 3 is the counterpart of Table 1 in which data are generated such that $\hat{\beta} = 0$ and $\hat{\sigma}^2 = 1$ and all other features are identical. Since $H_{\approx}$ is true the posterior predictive p-value should increases with N. Furthermore, since the data generated are more in agreement with $\beta \sim \mathcal{N}(0, .01)$ than with $\beta \sim \mathcal{N}(0, 1000000)$, the difference between both DIC's should increases with N. As can be seen, what should happen does not happen. Again this is caused by the fact that for increasing N the influence of the small variance prior on the posterior distribution disappears.

The phenomenon that the data dominate the prior is not restricted to the simple regression model. This implies that more elaborate structural equation models with corresponding generalization of $H_{\approx}$ can also not be evaluated using $p_{\approx}$ or $DIC_{\approx}$. What will result for increasing N is an evaluation of the fit of the model at hand as if uninformative instead of informative priors were used for the target parameters. In the next section this will be illustrated using a factor model with cross-loadings.

Performance of $p_{\approx}$ and $DIC_{\approx}$ for the Evaluation of Small Variance Priors for Cross Loadings

In this section two situations will be discussed: the evaluation of small variance priors for cross loadings in a two factor model that is identified, and the evaluation in a two factor model that is over identified. Using BIEMS data were generated such that in a two factor model with $p = 1, ..., 6$ indicators $z_1, ..., z_6$ the maximum likelihood estimates of the loadings are $\hat{\lambda}_{11} = 7, \hat{\lambda} = .7, \hat{\lambda}_{13} = .7, \hat{\lambda}_{14} = -.4, \hat{\lambda}_{15} = 0, \hat{\lambda}_{16} = .4$ for the first factor f_1 and $\hat{\lambda}_{21} = -.4, \hat{\lambda}_{22} = 0, \hat{\lambda}_{23} = .4, \hat{\lambda}_{24} = .7, \hat{\lambda}_{25} = .7, \hat{\lambda}_{26} = .7$ for the second factor f_2. The sample correlation between both factors equals 0, and the sample mean and variance are 0 and 1, respectively, for both factors. Finally, the residual variances of each indicator are $\hat{\tau}_1 = .35, \hat{\tau}_2 = .51, \hat{\tau}_3 = .35, \hat{\tau}_4 = .35, \hat{\tau}_5 = .51, \hat{\tau}_6 = .35$, that is, each indicator has a sample mean of 0 and sample variance of 1.

The following factor model was used to analyze these data for sample sizes of 50, 100, 500, 1000, and 5000, respectively:

$$z_{pi} = \lambda_{1p} f_{1i} + \lambda_{2p} f_{2i} + \epsilon_p \text{ with } \epsilon_p \sim \mathcal{N}(0.\tau_p) \text{ for } i = 1, ..., N, \tag{14}$$

and

$$\begin{bmatrix} f_{1i} \\ f_{2i} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \text{ for } i = 1, ..., N. \tag{15}$$

As can be seen, the model is not identified because in addition to fixing both factor variances to 1 two more constraints are needed. A standard manner to achieve this is by fixing one cross-loading to 0 for each factor. Muthen and Asparouhov (2012) identify the model using small variance priors for the cross-loadings (the smaller loadings in our example), that is, $\lambda_{14}, \lambda_{15}, \lambda_{16}, \lambda_{21}, \lambda_{22}, \lambda_{23}$. We will use two variations of their approach: the first renders a factor model that is indentified (I),

$$H_{\approx.I} : h(\lambda_{fp}) \sim \mathcal{N}(0, .01) \text{ for } fp = 15, 22; \tag{16}$$

the second renders a factor model that is over-identified (IO),

$$H_{\approx.OI} : h(\lambda_{fp}) \sim \mathcal{N}(0, .01) \text{ for } fp = 14, 15, 16, 21, 22, 23. \tag{17}$$

Standard uninformative priors are used for the other model parameters.

It is important to note, that it is not possible to test the small variance priors used to identify the model. When the model is identified by fixing one loading for each factor at a specific value, the fit of the model is unaffected by the value chosen, because the data do *not* contain information about this value. The same holds for identification by means of a small variance prior for one loading of each factor. As can be seen in the second column of Table 4 (obtained using Mplus using fbiterations=100000), independent of the sample, size $p_{\approx,I}$ indicates that the fit of the model to the data is good. This is as it should be, because the data are generated to perfectly fit the model used. However, if additional small variance priors are specified as in the over-identified model, the data can be used to test the appropriateness of these priors. As can be seen above, for $fp = 14, 16, 21, 23$, $\hat{\lambda}_{fp}$ is rather different from the mean of the corresponding small variance prior. Therefore, with increasing sample sizes, the evidence against $H_{\approx.OI}$ should increase. As can be seen in Table 4, this does not happen: $p_{\approx,OI}$ increases with increasing sample sizes and the differences between $DIC_{\approx,I}$ and $DIC_{\approx,OI}$ also decrease instead of increasing in favor of $DIC_{\approx,I}$.

The main conclusion obtained from the results presented in this and the previous section is that the posterior predictive p-value and the DIC can not be used for the evaluation of small variance priors because their behavior is inconsistent. One of the reviewers wondered whether modifications are feasible such that the posterior predictive p-value and the DIC can be used for the evaluation of small variance priors. The suggestion to reduce the size of the prior variance with increasing sample size might work if a clear cut procedure how to achieve this can be developed. However, we do not endorse this because we want to choose the prior variance such that on account of substantive reasons and irrespective of the sample size it represents what constitute relevant and irrelevant differences from zero. The suggestion to replace the standard DIC based on samples from the posterior distribution by a "prior" DIC based on samples from the prior distribution was explored by Van de Schoot et al. (2011). They did not address small-variance priors but other types of informative priors and showed that a prior DIC is also not suited for the evaluation of informative prior distributions. In the next section the prior-posterior predictive p-value will be introduced. It will be shown that this p-value is suited for the evaluation of small variance priors.

Evaluating Small Variance Priors Using The Prior-Posterior Predictive P-value

The Prior-Posterior Predictive P-value

To evaluate the fit of $H_0 : \beta = 0$ in the context of the simple regression model, the following posterior predictive p-value is used (Scheines, Hoijtink, and Boomsma, 1999):

$$p_0 = P(D_{rep} > D|\boldsymbol{y}, \boldsymbol{x}, \beta = 0) =$$

$$\int_{\alpha,\sigma^2} P(D_{rep} > D|\alpha, \sigma^2, \beta = 0) g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, \beta = 0) d\alpha, \sigma^2. \tag{18}$$

To test $H_{\approx} : \beta \sim \mathcal{N}(0, \tau^2)$, p_0 from Equation 18 can be modified into the prior-posterior predictive p-value

$$p_{pripos} = P(D_{rep} > D|\boldsymbol{y}, \boldsymbol{x}, h(\beta))$$

$$\int_{\beta} \int_{\alpha,\sigma^2} P(D_{rep} > D|\alpha, \sigma^2, \beta) h(\beta) g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, h(\beta)) d\alpha, \sigma^2 d\beta, \tag{19}$$

where

$$g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, h(\beta)) = \int_{\beta} g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, \beta) h(\beta) d\beta, \tag{20}$$

y_{rep} is simulated using $\theta = [\alpha, \sigma^2, \beta]$, and D_{rep} and D are based on

$$\Sigma^{rep}(\theta) = \begin{bmatrix} s^2_{\boldsymbol{y}_{rep}} & 0 \\ 0 & s^2_{\boldsymbol{x}} \end{bmatrix}, \tag{21}$$

and

$$\Sigma(\theta) = \begin{bmatrix} s^2_{\boldsymbol{y}} & 0 \\ 0 & s^2_{\boldsymbol{x}} \end{bmatrix}, \tag{22}$$

respectively, that is, independent of θ which makes D_{rep} and D test statistics instead of discrepancy measures. Note that, for increasing N the prior-posterior predictive p-value from Equation 19 does not behave as if an uninformative prior distribution for β was specified because the effect of $h(\beta)$ on the computation of p_{pripos} is direct and not mediated through the posterior distribution as was the case for the posterior predictive p-value.

Performance, Null-Distribution and Power of p_{pripos}

In this section three aspects of the performance of p_{pripos} will be illustrated. First of all, it will be shown that it does not suffer from the same drawbacks as the posterior predictive p-value and the DIC. Secondly, it will be shown that the null-distribution of p_{pripos} is approximately uniform. Thirdly, it will be shown using small variance priors for the difference between two means and a correlation, that larger sample sizes are needed to obtain a power of .80 than when the corresponding classical null-hypotheses are tested using a classical p-value.

In Appendix A, the algorithm used to compute p_{pripos} is elaborated. This algorithm is used to illustrate the performance of p_{pripos} in Table 5. Data are generated using BIEMS such that the sample mean and variance of $\boldsymbol{x}$ are 0 and 1, respectively, $\hat{\alpha} = 0$, for various values of $\hat{\beta}$ and $\hat{\sigma}^2$ chosen such that the variance of $\boldsymbol{y}$ equals 1. The results in Table 5 highlight four features of p_{pripos}:

1. When $\hat{\beta} = 0$, $p_{pripos} = 1$, independent of the sample size. This is analogous to the classical p-value corresponding to $H_0 : \beta = 0$ when $\hat{\beta} = 0$, that is, if the sample estimate is identical to the null-value, the resulting p-value equals 1.

2. When $\hat{\beta}$ increases while keeping the variance of $\boldsymbol{y}$ fixed at 1 (implying the the proportion of variance explained increases with $\hat{\beta}$), p_{pripos} decreases, that is, the larger the evidence against $H_{\approx}$ the smaller p_{pripos}. This is the desired behavior of the p-value and can be observed for each sample size.

3. For $\hat{\beta} > 0$ the p_{pripos} decreases with increasing N. This too is the desired behavior of the p-value.

4. Comparing the one but last column of Table 5 with the third column of Table 1 (reproduced in the last column of Table 5) it can be observed for $N = 25$ and $N = 100$ that $p_{pripost}$ is smaller than $p_{\approx}$ and thus is more powerful in rejecting $H_{\approx}$.

Based on the results in Table 5 it can be concluded that the basic behavior of p_{pripos} is adequate. What remains to be determined, are guidelines for the interpretation of the size of $p_{pripost}$. In Figure 1 a null distribution of p_{pripos} is displayed. This null-distribution was obtained after executing the following four steps 100000 times:

1. Sample x_i from $\mathcal{N}(0, 1)$ for $i = 1, ..., N$,

2. Sample β from $\mathcal{N}(0, .01)$ and set $\alpha = 0$, $\sigma^2 = 1$,

3. Simulate y_i from $\mathcal{N}(\alpha + \beta x_i, \sigma^2)$ for $i = 1, ..., N$,

4. Compute p_{pripos},

and present the 100000 resulting p-values in a histogram. As can be seen in Figure 1, the deviations from uniformity (expected frequency of each bar equal to 5000) are rather small and never more than 150. Furthermore, the probability that p_{pripos} is smaller than .05 is about .0512. It seems that it is reasonable to interpret p_{pripos} as a classical p-value that is uniform under the null, that is, compare p_{pripos} to a type I error level of .05.

Cohen (1992) presents sample size calculations for testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$, where μ_1 and μ_2 denote the means in groups 1 and 2, respectively. As can be seen in the top panel of Table 6, if the type I error is controlled at .05, to achieve a power of .80, the number of persons per group should be 393, 64, and 26 for small, medium and large effects $d = (\mu_1 - \mu_2)/\sigma$, respectively. Using $x_i = 1$ for $i = 1, ..., .5N$ and $x_i = 0$ for $i = .5N + 1, ..., N$ implies that $\beta = \mu_1 - \mu_2$ and $\alpha = \mu_2$. Repeatedly sampling data from a population in which $\mu_1 = d$, $\mu_2 = 0$ and $\sigma = 1$ allows the computation of the sample size for testing $H_{\approx} : \beta \sim \mathcal{N}(0, .01)$ such that the power is about .80 if the type I error is controlled at .05. As can be seen in Table 6, compared to testing the traditional null-hypothesis larger sample sizes are needed to test small variance priors with the same power. This is not surprising, because it is easier to reject a precise than a less precise hypothesis. More specifically, it can be seen that unrealistically large sample sizes are needed to detect an effect $d = .2$. This is not surprising, because $d = .2$ is rather realistic if $\beta \sim \mathcal{N}(0, .01)$. Furthermore, it can be seen that somewhat larger sample sizes are need to detect $d = .5$, and that similar sample sizes are needed to detect $d = .8$.

When testing $H_{\approx}$, it is important that the prior variance of β is tailored to the scale of the data. In the power analysis presented, the population parameters were chosen such that $\beta = d$. Consequently, the prior variance of .01 implies that under $H_{\approx}$ values of $\beta = d$ between -.2 and +.2 are rather likely. Stated otherwise, an observed effect size has to be larger than “small” to cast doubt on $H_{\approx}$. As will be elaborated in the next section, when analyzing empirical data with the approach proposed in this paper, tailoring the prior variance of β to the scale of the data at hand is an important step.

In the bottom panel of Table 6, Cohen’s (1992) sample sizes for testing $H_0 : \rho = 0$ are displayed. Repeatedly sampling data from a population in which

$$\begin{bmatrix} y_i \\ x_i \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \text{ for } i = 1, ..., N, \tag{23}$$

implies that $\alpha = 0$, $\beta = \rho$, and $\sigma^2 = 1 - \rho^2$. With this setup a prior variance of .01 for β implies that under $H_{\approx}$ values of $\beta = d$ between -.2 and +.2 are rather likely. Stated otherwise, observed correlations have to be approaching “medium” to cast doubt on $H_{\approx}$.

As can be seen, the results obtained are analogous to the results obtained in the top panel of Table 6. The main difference is that the sample size needed to detect a medium effect is a about twice larger than the corresponding "classical" sample size. All in all, it can be concluded that for effect sizes that are "unreasonable" under $H_{\approx}$ the sample sizes needed to obtain a power of .80 are larger than the sample sizes needed in the classical setting. For effect sizes that are "reasonable" under $H_{\approx}$ the sample sizes needed are unrealistically large. These features are appropriate if the goal is to evaluate $H_{\approx}$ instead of a classical null-hypothesis.

Examples

In this section two applications of small variance priors will be presented. The first concerns testing whether the difference between two means is relevantly different from zero. The second concerns testing whether a correlation is relevantly different from zero.

Testing Cohen's d

Henderson, de Liver, and Gollwitzer (2008) present research with respect to the relation between mind-set and attitude strength with respect to the statement that "a list of sex offenders should be made public". Attitude strength is measured on a scale ranging from -3 to +3. The higher the score, the smaller the attitude strength. Two of their mind-set groups are: "one-sided", in which participants are primed to rely on their own experiences which could be against or in favor of the statement; and "two-sided", in which participants are primed that both perspectives are important. In Table 7 descriptives and p-values obtained for the comparison of both group with respect to attitude strength are presented.

Using Students' t-test to evaluate $H_0 : \mu_1 = \mu_2$ rendered a classical p-value of .021, that is, using the conventional type I error equal to .05 implies that H_0 has to be rejected. An interesting question is whether the null-hypothesis can also be rejected if it specifies that the difference between both means is irrelevantly different from zero. This can be investigated using the following three step procedure:

1. Specify the prior variance corresponding to an irrelevant difference. Note that the relevance of a difference depends on σ^2. Here a prior variance of $\tau = .01$ is considered to represent irrelevant differences if $\sigma^2 = 1$.

2. Tailor the prior variance to the scale of the data at hand. Use $\hat{\sigma} = 1.855$ to compute the rescaled prior variance $\tau = (.1 \times 1.855)^2$.

3. Use the same set-up as in the sample size calculation in the previous section and τ as determined in the previous step to compute p_{pripos} for $H_{\approx} : \beta \sim \mathcal{N}(0, .0344)$.

These three steps result in $p_{pripos} = .024$ which is only marginally larger than the corresponding classical p-value. This implies that the difference between means of attitude strength in the one and two sided groups are not only different from zero, but also relevantly different from zero.

Testing a Correlation

Dolan, Oort, Stoel, and Wicherts (2009) obtained data for the BIG5 personality inventory from 500 students. These data are incorporated in the package JASP (https://jasp-stats.org). Here the scores on the variables “openness” and “agreeableness” will be used to test whether the correlation between both variables is irrelevantly different from zero. Descriptives for both variables are presented in Table 8. As can be seen, the classical p-value computed for $H_0 : \rho = 0$ equals .00, that is, using the convential type I error level of .05, it can be concluded that H_0 has to be rejected. The following three steps can be used to test whether the correlation is irrelevantly different from zero:

1. Specify the prior variance corresponding to an irrelevant difference. Note that the relevance of a difference depends on the scale of $\boldsymbol{y}$ and $\boldsymbol{x}$. Here a prior variance of

$\tau = .01$ is considered to represent irrelavant differences if both variables are standardized because in that situation $\beta = \rho$.

2. Tailor the prior variance to the scale of the data at hand. Use $s^2_{\boldsymbol{y}} = .34^2$ and $s^2_{\boldsymbol{x}} = .35^2$ to compute the rescaled prior variance $\tau = (.1 \times .34/.35)^2$.

3. Use the same set-up as in the sample size calculation in the previous section and τ as determined in the previous step to compute p_{pripos} for $H_{\approx} : \beta \sim \mathcal{N}(0, .0094)$.

These three steps result in $p_{pripos} = .58$ which is larger than .05. It can therefore be concluded that although ρ is different from zero, it is not relevantly different from zero.

Discussion

The title of Cohen's (1994) paper " The earth is round, p<.05" nicely stresses that small and irrelevant differences from a null value may result in a rejection of the corresponding null-hypothesis. This problem is addressed by Muthen and Asparouhov (2012) who propose so-called small variance priors for the parameters of interest that represent irrelevant differences from zero. This paper addressed testing the hypothesis that parameters are irrelevantly different from zero. Using evaluations based on a simple regression model four conclusions were obtained:

1. The posterior predictive p-value and the DIC are inconsistent. If $H_{\approx}$ is not true and the sample size increases both express *decreasing* evidence against $H_{\approx}$. If $H_{\approx}$ is true and the sample size increases both express *increasing* evidence against $H_{\approx}$.

2. The prior-posterior predictive p-value does show the desired behavior, both if the hypothesis specifying irrelevant differences is false and true. It is more powerful than the posterior predictive p-value and has a null-distribution that is approximately uniform. It appears to be an excellent p-value for the evaluation of hypotheses specifying small variance priors for the parameters of interest.

3. Because the traditional null-hypothesis is more specific than the hypothesis specifying irrelevant differences, evaluations of the latter are less powerful. Compared to the traditional null-hypothesis larger sample sizes are needed to achieve the same power.

4. The prior-posterior predictive p-value can be used for the evaluation of hypotheses specifying irrelevant differences from a null-value for the differences between two means and correlations.

Although the evaluations of the prior-posterior predictive p-value presented in this paper were limited to the context of a simple regression model, the main conclusions generalize to structural equation models. There too the influence of the small variance prior will not disappear with increasing sample sizes, because the prior-posterior predictive p-value is based on data sets replicated using the prior distribution of the parameters of interest and the posterior distribution (conditional on the aforementioned prior distribution) of the remaining parameters. Of course, more elaborate evaluations of the prior-posterior predictive p-value in the context of structural equation models are needed. No doubt these will be made as soon as the authors of one or more of the software packages referred to in this paper implement it, because testing whether a model approximately fits the data is a much better idea than testing whether a model exactly fits the data.

It is important to note that evaluation of small variance priors using the prior-posterior predictive p-value is one alternative for classical null-hypothesis significance testing. The interested reader is referred to Morey and Rouder (2011) and Hoijtink (2012, pp. 8-10) who replace the classical null hypothesis by an interval null hypothesis and evaluate it by means of the Bayes factor (Kass and Raftery, 1995). The main difference with the approach presented in this paper is that the dichotomous reject/do not reject decision based on the comparison of the p-value with a pre-specified alpha level is replaced by the relative support in the data for the hypotheses entertained as expressed by the Bayes factor. Another alternative is to summarize the information, e.g. with respect to the difference between two means, in a confidence or credible interval (Cumming, 2012) and to provide a qualitative interpretation. However, both approaches are currently essentially limited to the evaluation of one quantity of interest (e.g., the difference between two means) and further research is needed to explore if and how they can be applied in more elaborate models like the factor model with cross-loadings.

References

Asparouhov, T., Muthen, B., and Morin, A.J.S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: comments on Stromeyer et al. *Journal of Management, 41*, 1561-1577. doi:10.1177/0149206315591075

Bujacz, A., Vitters, J., Huta, V., and Kaozmarek, L. D. (2014). Measuring hedonia and eudaimonia as motives for activities: cross-national investigation through traditional and Bayesian structural equation modeling. *Frontiers in Psychology, 5*, 984. doi:10.3389/fpsyg.2014.00984

Chiorri, C., Day, T., and Malmberg, L.-E. (2014). An approximate measurement invariance approach to within-couple relationship quality. *Frontiers in Psychology, 5*, 983. doi:10.3389/fpsyg.2014.00983

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., and Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5:982.* doi:10.3389/fpsyg.2014.00982

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2012). *Understanding the New Statistics. Effect Sizes, Confidence Intervals, and Meta Analysis.* New York: Routledge.

Dolan, C.V., Oort. F.J., Stoel, R.D., and Wicherts, J.M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling: A Multidiciplinary Journal, 16,* 295-314. doi:10.1080/10705510902751416

Henderson, M.D., de Liver, Y., and Gollwitzer, P.M. (2008). The effects of an implemental mind-set on attitude strenght. *Journal of Personality and Social Psychology, 94,* 396-411. doi:10.1037/0022-3514.94.3.396

Hoijtink, H. (2012). *Informative Hypotheses. Theory and Practice for Behavioral and Social Scientists.* Boca Raton: Chapmann and Hall/CRC.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*, 773-795. DOI: 10.1080/01621459.1995.10476572

Kelcey, B., McGinn, D., and Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology, 5*, 1469. doi:10.3389/fpsyg.2014.01469

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS: A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing,* 10, 325–337. doi:10.1023/A:1008929526011

Maccallum, R.C., Edwards, M.C., and Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods, 17,* 340-345. doi:10.1037/a0027131

Meng, X-L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22,* 1142-1160. doi:10.1214/aos/1176325622

Merkle, E.C. and Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. http://arxiv.org/abs/1511.05604

Morey, R. D. and Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods, 16,* 406-419. http://dx.doi.org/10.1037/a0024377

Mulder, J., Hoijtink, H., and de Leeuw, C. (2012). BIEMS, a Fortran90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software, 46,* 2. http://dx.doi.org/10.18637/jss.v046.i02

Muthen,B.O. and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods, 17,* 313-335. http://dx.doi.org/10.1037/a0026802

Muthen, L.K. and Muthen, B.O. (1998-2015). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthen & Muthen. Retrieved from: https://www.statmodel.com/

Rindskopf, D. (2012). Next steps in Bayesian structural equation models: Comments on, variations of, and extensions to Muthen and Asparouhov (2012). *Psychological Methods, 17,* 336-339. http://dx.doi.org/10.1037/a0027130

Scheines, R., Hoijtink, H., and Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64,* 37-52. doi:10.1007/BF02294318

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 4*, 583-639. doi:10.1111/1467-9868.00353

Stromeyer, W.R., Millar, J.W., Sriramachandramurthy, R., and DeMartino, R. (2015). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations for management research. *Journal of Management, 41*, 491-520. doi:10.1177/0149206314551962

Van de Schoot, R., Hoijtink, H., Romeijn, J-W, and Brugman, D. (2011). A prior predictive loss function for the evaluation of inequality constrained hypotheses. *Journal of Mathematical Psychology, 56*, 13-23. doi:10.1016/j.jmp.2011.10.001.

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. and Muthen, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4:770.* doi: 10.3389/fpsyg.2013.00770.

Zercher, F., Schmidt, P., Cieciuch, J., and Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: exact versus approximate measurement equivalence. *Frontiers in Psychology, 6*, 733. doi:10.3389/fpsyg.2015.00733

Appendix A: The Algorithm Used to Compute p_{pripos}

To compute p_{pripos} from Equation 19, first of all a five step Markov chain Monte Carlo algorithms is used to sample from $h(\beta)g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, h(\beta))$.

Note, that, $g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, h(\beta)) \approx 1/T \sum_{t=1}^{T} g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, \beta_t)$, where β_t for $t = 1, ..., T$ is sampled from $h(\beta)$. Sampling α, σ^2 from $g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, \beta_t)$ for various values of β_t renders a sample from $g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, h(\beta))$.

1. Set $\sigma_0^2 = 1$, that is, in iteration $t = 0$ an initial value is assigned to σ^2.

2. Execute Steps 3 through 5 for t=1,...,10000 iterations.

3. Sample β_t from $h(\beta)$. Set $\sigma_{0,t-1}^2 = \sigma_{t-1}^2$, that is, in iteration $u = 0$ an initial value is assigned.

4. For $u = 1, ..., U$ execute the following two steps, that is, a Gibbs sampler to sample from $g(\alpha, \sigma^2|\boldsymbol{y}, \boldsymbol{x}, \beta_t)$:

(a) Sample $\alpha_{u,t-1}$ from $g(\alpha|\boldsymbol{y},\boldsymbol{x},\beta_t,\sigma^2_{u-1,t-1}) = \mathcal{N}(m,s^2)$ where $m = \sum_{i=1}^N (y_i - \beta_t x_i)/N$ and $s^2 = \sigma^2_{u-1,t-1}/N$.

(b) Sample $\sigma^2_{u,t-1}$ from $g(\sigma^2|\boldsymbol{y},\boldsymbol{x},\beta_t,\alpha_{u,t-1}) = \Gamma^{-1}(a,b)$ where $a = N/2 - 1$ and $b = .5\sum_{i=1}^N (y_i - \alpha_{t-1} - \beta_t x_i)^2$.

5. Set $\alpha_t = \alpha_{U,t-1}$ and $\sigma^2_t = \sigma^2_{U,t-1}$.

Subsequently, Equation 19 is approximated by

$$p_{pripos} \approx \sum_{t=1}^{10000} P(D_{rep,t} > D|\alpha_t,\sigma^2_t,\beta_t), \tag{24}$$

where $y_{rep,t}$ is generated using $\alpha_t,\sigma^2_t,\beta_t$ and $D_{rep,t}$ is a function of $y_{rep,t}$ (cf. Equation 21).

With $U = 1000$ it is clear that in Step 4 $g(\alpha,\sigma^2|\boldsymbol{y},\boldsymbol{x},\beta_t)$ is sampled from because the dependence on the initial value σ^2_{t-1} will disappear. However, the resulting sample can not be distinguished from the one obtained using $U = 1$. This makes the sampler for $h(\beta)g(\alpha,\sigma^2|\boldsymbol{y},\boldsymbol{x},h(\beta))$ as efficient as samplers that address $g(\beta,\alpha,\sigma^2|\boldsymbol{y},\boldsymbol{x})$. Data with $N = 25$ are generated using BIEMS such that $\hat{\alpha} = 0$, $\hat{\beta} = .3$, $\hat{\sigma^2} = .91$, and the sample mean and variance of $\boldsymbol{x}$ are 0 and 1, respectively. As can be seen in Table 9, using the algorithm described above to sample from $h(\beta)g(\alpha,\sigma^2|\boldsymbol{y},\boldsymbol{x},h(\beta))$ with $U = 1$ and $U = 1000$ rendered distributions for β, α, and σ^2, that are virtually identical.

For $U = 1$ the sampler is almost a traditional Gibbs sampler, except that β is sampled from its prior instead of its posterior distribution. As has been shown, for the simple model at hand a burn-in period is not needed to sample the nuisance parameters (α,σ^2) from their posterior distribution conditional on β. Although it is expected that the same holds for models involving more than two nuisance parameters, users of the prior-posterior predictive p-value are well-advised to explore whether a burn-in period is needed before choosing U.

Table 1

Performance of $p_{\approx}$ and $DIC_{\approx}$ for $H_{\approx}$ is not True

N	$p_{\approx}$	$p_{\approx}$	$DIC_{\approx}$	$DIC_{\approx}$	$q_{\approx}$	$q_{\approx}$
	$\tau^2 = 1000000$	$\tau^2 = .01$	$\tau^2 = 1000000$	$\tau^2 = .01$	$\tau^2 = 1000000$	$\tau^2 = .01$
25	.46	.01	59.92	69.96	2.72	1.80
100	.50	.02	220.44	232.76	2.91	2.51
1000	.50	.31	2150.38	2152.59	2.97	2.92
10000	.49	.47	21450.39	21450.64	3.03	3.03

Table 2

The Counterpart of Table 1 in which Equation 1 is Extended with $x_i \sim \mathcal{N}(\mu, \sigma^2_\mu)$

N	$p_\approx$	$p_\approx$	$DIC_\approx$	$DIC_\approx$	$q_\approx$	$q_\approx$
	$\tau^2 = 1000000$	$\tau^2 = .01$	$\tau^2 = 1000000$	$\tau^2 = .01$	$\tau^2 = 1000000$	$\tau^2 = .01$
25	.45	.02	134.92	144.81	4.46	3.46
100	.49	.03	508.43	520.47	4.90	4.41
1000	.50	.34	4992.24	4994.37	4.95	4.88
10000	.49	.48	49833.14	49833.34	5.01	4.99

Table 3

Performance of $p_{\approx}$ and $DIC_{\approx}$ for $H_{\approx}$ is True

N	$p_{\approx}$	$p_{\approx}$	$DIC_{\approx}$	$DIC_{\approx}$	$q_{\approx}$	$q_{\approx}$
	$\tau^2 = 1000000$	$\tau^2 = .01$	$\tau^2 = 1000000$	$\tau^2 = .01$	$\tau^2 = 1000000$	$\tau^2 = .01$
25	.46	.59	77.26	75.45	2.72	1.94
100	.50	.56	289.79	288.75	2.91	2.40
1000	.50	.51	2843.83	2843.65	2.97	2.88
10000	.50	.50	28384.82	28384.80	3.03	3.02

Table 4

Performance $p_{\approx}$ and $DIC_{\approx}$ in the Two Factor Model (I denotes identified, OI denotes over identified)

N	$p_{\approx,I}$	$p_{\approx,OI}$	$DIC_{\approx,I}$	$DIC_{\approx,OI}$	$q_{\approx,I}$	$q_{\approx,OI}$
50	.54	.08	796.48	813.09	21.39	19.38
100	.61	.16	1542.77	1557.24	22.37	21.34
500	.65	.45	7522.53	7527.78	22.57	22.76
1000	.65	.55	14998.92	15002.12	22.47	22.94
5000	.66	.63	74810.46	74812.79	21.43	23.16

Table 5

Performance of p_{pripos}

N			$\hat{\beta}/\hat{\sigma}^2$			$p_{\approx}$
	0/1	.1/.99	.2/.96	.3/.91	.707/.50	.707/.50
25	1	.66	.38	.17	.00	.01
50	1	.56	.24	.08	.00	
100	1	.47	.15	.03	.00	.31

Table 6

Sample Size Calculations for Type I Error Probability .05 and Power .80

	small	medium	large
d	.20	.50	.80
N per group for $H_0 : \mu_1 = \mu_2$	393	64	26
N per group for $H_{\approx} : \beta \sim \mathcal{N}(0, .01)$	>20000	80	28
ρ	.10	.30	.50
N for $H_0 : \rho = 0$	783	85	28
N for $H_{\approx} : \beta \sim \mathcal{N}(0, .01)$	>20000	169	33

Table 7

Descriptives and P-values for the Henderson et al. (2008) Data ($\hat{d} = .89$)

		Attitude Strength	
group	N	mean	sd
one-sided	15	.16	1.85
two-sided	15	1.82	1.86
$H_0 : \mu_1 = \mu_2$	classical $p = .021$		
$H_{\approx} : \beta \sim \mathcal{N}(0, .0344)$	$p_{pripos} = .024$		

Table 8

Descriptives and P-values for the Dolan et al. (2009) Data (N=500, $\hat{\rho} = .16$)

variable	mean	sd
openness	3.59	.34
agreeableness	3.42	.35
$H_0 : \rho = 0$	classical $p = .00$	
$H_{\approx} : \beta \sim \mathcal{N}(0, .0094)$	$p_{pripos} = .58$	

Table 9

Identical Samples for β, α, and σ^2 using $U = 1$ and $U = 1000$

percentile	α	α	σ^2	σ^2	β	β
	$U = 1$	$U = 1000$	$U = 1$	$U = 1000$	$U = 1$	$U = 1000$
.05	-.37	-.38	.76	.77	-.16	-.17
.10	-.29	-.29	.84	.85	-.13	-.13
.25	-.14	-.15	1.00	1.01	-.07	-.07
.50	.00	.00	1.24	1.23	.00	.00
.75	.15	.15	1.55	1.54	.07	.07
.90	.29	.29	1.92	1.92	.13	.13
.95	.38	.38	2.20	2.21	.16	.16

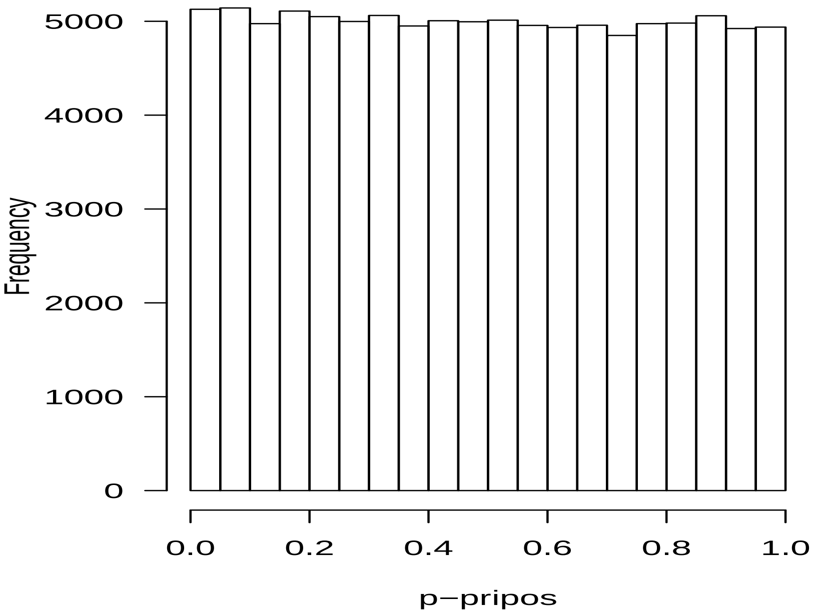

Figure 1. The Null Distribution of p_{pripos} for $N = 25$, $\alpha = 0$, $\sigma^2 = 1$, and $x_i \sim \mathcal{N}(0, 1)$