Supplement to "Statistical Methods for Replicability Assessment"

Kenneth Hung[*] William Fithian[†]

March 13, 2019

1 t-tests and ANOVAs

Both t-tests and $F(1, \cdot)$ ANOVAs, whether they are within-subject or between-subject, one-way or multi-way, can be thought of as testing some contrast δ of the true cell means.

We can construct unbiased estimators $\hat{\delta}$ and $\hat{\sigma}^2$ for the contrast and the common variance from the homoscedasticity assumption. The standard error can be written as $\frac{\sigma}{k}$ where σ^2 is the true variance and k is some known constant determined by the cell sizes. We can now test the null hypothesis $\delta = 0$ using the t-statistic $T = k\hat{\delta}/\hat{\sigma}$. If there is an effect, i.e. $\delta \neq 0$, then T follows the noncentral t-distribution

$$T \sim t_{df}\left(k\frac{\delta}{\sigma}\right).$$

We take the effect size θ to be δ/σ.

For notation, we will use subscripts to indicate the cell. We will use μ_i (or, μ_{ij} or μ_{ijk}) to denote the population mean in each cell. For the marginals, we use $*$ to indicated dimensions averaged out. For example, in a 2×3 ANOVA,

$$\mu_{1*} = \frac{\mu_{11} + \mu_{12} + \mu_{13}}{3}.$$

For designs with between-subject factor(s), the subject can be naturally grouped, we will use I_i (or I_{ij}) to indicate these groups. The number of subjects in each is n_i (or n_{ij}) respectively. For example, if there are two between-subject factors, each with two levels, we have four groups I_{11}, I_{12}, I_{21} and I_{22}, and their sizes are n_{11}, n_{12}, n_{21} and n_{22} respectively. Finally, when a subject is specified and their between-subject factor is clear from the context, we may omit the index as $\cdot$.

Study 1: Roelofs (2008) The study has a $2 \times 2 \times 2$ within-subject design and we are interested in the two-way interaction between the second and third factors. In other words, we want to test if

$$\delta = \mu_{*11} - \mu_{*12} - \mu_{*21} + \mu_{*22}$$

[*] Department of Mathematics, University of California, Berkeley
[†] Department of Statistics, University of California, Berkeley

is zero.

We can test it by finding an unbiased estimator of this, and divide it by its standard error estimate to give a t-statistic. A natural estimate is

$$\hat{\delta} = \frac{1}{n}\sum \lambda' X^{(i)},$$

where $\lambda' = (\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2})$ and $X_{(i)}$ is the observation

$$(X_{111}, X_{112}, X_{121}, X_{122}, X_{211}, X_{212}, X_{221}, X_{222})'$$

for individual i. We can view $\lambda' X^{(i)}$ as the observations and perform a one-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = \delta/\sigma$ and σ^2 is the variance of $\lambda' X^{(i)}$. We have $n_O = 14$ and $n_R = 29$.

Study 2: Morris and Still (2008) The study is a within-subject one-way ANOVA. In other words, we want to test if

$$\delta = \mu_1 - \mu_2$$

is zero. A natural estimate for δ is

$$\hat{\delta} = \frac{1}{n}\sum \lambda' X^{(i)},$$

where $\lambda' = (1, -1)$ and $X^{(i)}$ is the observation $(X_1, X_2)'$ for individual i. We can now view $\lambda' X^{(i)}$ as the observations and perform a one-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = \lambda'(\mu_1, \mu_2)/\sigma$ and σ^2 is the variance of the contrast. We have $n_O = n_R = 24$.

Study 3: Liefooghe et al. (2008) The RP:P dataset describes this as a multivariate ANOVA, but upon inspection of its code (`https://osf.io/69b27/`) we recognize the test of interest is in fact a paired t-test of two components of the multivariate observation. We can consider the differences as the observations. Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = (\mu_1 - \mu_2)/\sigma$ and σ^2 is the variance of the difference. We have $n_O = 25$ and $n_R = 32$.

Study 4: Storm et al. (2008) The study is a 2×2 mixed design with the first factor being within-subject and the second factor being between-subject. We are interested in the interaction. In other words, we want to test if

$$\delta = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$$

is zero. A natural estimate for δ is

$$\hat{\delta} = \frac{1}{n_1}\sum_{i \in I_1} \lambda' X^{(i)} - \frac{1}{n_2}\sum_{i \in I_2} \lambda' X^{(i)},$$

where $\lambda' = (1, -1)$ and $X^{(i)}$ is the observation $(X_{1\cdot}, X_{2\cdot})'$ for subject i. We can view $\lambda' X_{(i)}$ as the observations and perform a two-sample t-test. We however cannot reanalyze the replication data to confirm that our test is equivalent as the data has since been removed (`https://osf.io/rj4u6/`).

Hence the test statistic T follows $T \sim t_{n_1+n_2-2}\left(\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1/2}\theta\right)$, where

$$\theta = \left(\lambda'(\mu_{11},\mu_{21}) - \lambda'(\mu_{12},\mu_{22})\right)/\sigma$$

and σ^2 is the variance of $\lambda' X_{(i)}$. We have $n_{1O}+n_{2O}=240$, $n_{1R}=136$ and $n_{2R}=134$ but n_{1O} and n_{2O} are unavailable.

Study 5: Mitchell et al. (2008) The study has a 2 × 2 within-subject design and we are interested in the main effect of the first factor. In other words, we want to test if

$$\delta = \mu_{1*} - \mu_{2*}.$$

is zero. A natural estimate for δ is

$$\hat{\delta} = \frac{1}{n}\sum \lambda' X^{(i)},$$

where $\lambda' = (\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2})$ and $X^{(i)}$ is the observation $(X_{11}, X_{12}, X_{21}, X_{22})'$. We can view $\lambda' X^{(i)}$ as the observations and perform a one-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = \delta/\sigma$ and σ^2 is the variance of $\lambda' X^{(i)}$. We have $n_O = 32$ and $n_R = 48$.

Study 6: Berry et al. (2008) The test of interest is a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 24$ and $n_R = 32$.

Study 7: Beaman et al. (2008) The test of interest is a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 100$ and $n_R = 15$.

Study 8: Dodson et al. (2008) The study has a 2 × 2 × 2 mixed design with the first two factors being within-subject and the third factor being between-subject. We are interested in the three-way interaction. In other words, we want to test if

$$\delta = \mu_{111} - \mu_{112} - \mu_{121} - \mu_{211} + \mu_{122} + \mu_{212} + \mu_{221} - \mu_{222}$$

is zero. A natural estimate of δ is

$$\hat{\delta} = \frac{1}{n_1}\sum_{i\in I_1} \lambda' X^{(i)} - \frac{1}{n_2}\sum_{i\in I_2} \lambda' X^{(i)}$$

where $\lambda' = (1, -1, -1, 1)$ and $X^{(i)}$ is the observation $(X_{11\cdot}, X_{12\cdot}, X_{21\cdot}, X_{22\cdot})'$ for subject i and $\cdot$ is the appropriate index. We can view $\lambda' X^{(i)}$ as the observations and perform a two-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n_1+n_2-2}\left(\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1/2}\theta\right)$, where $\theta = \delta/\sigma$ and σ^2 is the variance of $\lambda' X^{(i)}$. We have $n_{1O} = 20$, $n_{2O} = 19$, $n_{1R} = 13$ and $n_{2R} = 20$.

Study 10: Ganor-Stern and Tzelgov (2008) The study has a $2 \times 3 \times 2$ mixed design with the first two factors being within-subject and the third factor being between-subject. We are interested in the main effect of the first factor. In other words, we want to test if

$$\delta = \mu_{1**} - \mu_{2**}$$

is zero. A natural estimate of δ is

$$\hat{\delta} = \frac{1}{n_1} \sum_{i \in I_1} \lambda' X^{(i)} + \frac{1}{n_2} \sum_{i \in I_2} \lambda' X^{(i)}$$

where $\lambda' = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, -\frac{1}{6}, -\frac{1}{6}, -\frac{1}{6})$ and $X^{(i)}$ is the observation $(X_{11\cdot}, X_{12\cdot}, X_{13\cdot}, X_{21\cdot}, X_{22\cdot}, X_{23\cdot})'$ for subject i. We can view $\lambda' X_{(i)}$ as the observations and assume that it has the same variance σ^2 across the groups. The variance of $\hat{\delta}$ is thus $(\frac{1}{n_1} + \frac{1}{n_2})\sigma^2$, and σ^2 can be estimated with the pooled variance. Hence we can test if $\delta = 0$ with a t-test with the null distribution

$$T = \frac{\hat{\delta}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} \hat{\sigma}}} \sim t_{n_1+n_2-2} \left(\left(\frac{1}{n_{11}} + \frac{1}{n_{22}} \right)^{-1/2} \theta \right)$$

where $\theta = \delta/\sigma$. We have reanalyzed the replication data and confirmed that our test is equivalent. We have $n_{1O} + n_{2O} = 32$, $n_{1R} = 16$ and $n_{2R} = 15$ but n_{1O} and n_{2O} are unavailable.

Study 11: Mirman and Magnuson (2008) The study has a 2×2 within-subject design and we are interested in the main effect of the first factor. This is the same setup as Study 5, and hence we define δ, θ and σ^2 similarly. We have reanalyzed the replication data and confirmed that our test is equivalent. We have $n_O = 22$ and $n_R = 30$.

Study 15: Schmidt and Besner (2008) The test of interest is a paired t-test, same as Study 3. We defined θ similarly, and have $n_O = 95$ and $n_R = 242$.

Study 19: Oberauer (2008) The study has a 3×2 mixed design with the first factor being within-subject and the second factor being between-subject. We are interested in the main effect of the second fator. In other words, we want to test if

$$\delta = \mu_{*1} - \mu_{*2}$$

is zero. A natural estimate of δ is

$$\hat{\delta} = \frac{1}{n_1} \sum_{i \in I_1} \lambda' X^{(i)} - \frac{1}{n_2} \sum_{i \in I_2} \lambda' X^{(i)}$$

where $\lambda' = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $X^{(i)} = (X_{1\cdot}, X_{2\cdot}, X_{3\cdot})'$ for subject i. We can view $\lambda' X^{(i)}$ as the observations and perform a two-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n_1+n_2-2} \left(\left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1/2} \theta \right)$, where $\theta = \delta/\sigma$ and σ^2 is the variance of $\lambda' X^{(i)}$. We have $n_{1O} = n_{2O} = 16$, $n_{1R} = 11$ and $n_{2R} = 10$.

Study 20: Sahakyan et al. (2008) The study has a 2×2 mixed design with the first factor being between-subject and second factor being within-subject. We are interested in the interaction. This is the same setup as Study 4 and thus we define δ, θ and σ^2 similarly. We have reanalyzed the replication data and confirmed that our test is equivalent. We have $n_{1O} + n_{2O} = 96$, $n_{1R} = 47$ and $n_{2R} = 61$ but n_{1O} and n_{2O} are unavailable.

Study 24: Bassok et al. (2008) The study has a 2×2 within-subject design. We are interested in the interaction. In other words, we want to test if

$$\delta = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$$

is zero. A natural estimate of δ is

$$\hat{\delta} = \frac{1}{n} \sum \lambda' X^{(i)}$$

where $\lambda' = (1, -1, -1, 1)$ and $X^{(i)}$ is the observation $(X_{11}, X_{12}, X_{21}, X_{22})'$ for subject i. We can view $\lambda' X^{(i)}$ as the observations and perform a one-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = \delta/\sigma$ and σ^2 is the variance of the contrast. We have $n_O = 153$ and $n_R = 49$.

Study 27: Yap et al. (2008) The study has a 2×2 within-subject design. We are interested in the interaction. This is the same setup as Study 24 and thus we define δ, θ and σ^2 similarly. We have reanalyzed the replication data and confirmed that our test is equivalent. We have $n_O = 32$ and $n_R = 71$.

Study 29: Turk-Browne et al. (2008) The test of interest is a one-sample t-test. We are testing if the observations has mean 0.5. In other words, we want to test

$$\delta = \mu - 0.5$$

is zero. Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = (\mu/0.5)/\sigma$ and σ^2 is the variance of individual observations. We have $n_O = 8$ and $n_R = 15$.

Study 32: White (2008) The study has a 2×3 within-subject design. We are interested in the main effect of the first factor. In other words, we want to test if

$$\delta = \mu_{1*} - \mu_{2*}$$

is zero. A natural estimate for δ is

$$\hat{\delta} = \frac{1}{n} \sum \lambda' X^{(i)},$$

where $\lambda' = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$ and $X^{(i)}$ is the observation $(X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{23})'$. We can view $\lambda' X^{(i)}$ as the observations and perform a one-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = \delta/\sigma$ and σ^2 is the variance of $\lambda' X^{(i)}$. We have $n_O = 37$ and $n_R = 38$.

Study 33: Farrell (2008) The test of interest is a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 40$ and $n_R = 40$.

Study 36: Pacton and Perruchet (2008) The study has a $2 \times 2 \times 2$ mixed design with the first two factors being between-subject and the third factor being within-subject. We are interested in the main effect of the third factor. In other words, we want to test if

$$\delta = \mu_{**1} - \mu_{**2}$$

is zero. A natural estimate for δ is

$$\hat{\delta} = \frac{1}{n_{11}} \sum_{i \in I_{11}} \lambda' X^{(i)} + \frac{1}{n_{12}} \sum_{i \in I_{12}} \lambda' X^{(i)} + \frac{1}{n_{21}} \sum_{i \in I_{21}} \lambda' X^{(i)} + \frac{1}{n_{22}} \sum_{i \in I_{22}} \lambda' X^{(i)}$$

where $\lambda' = (\frac{1}{4}, -\frac{1}{4})$ and $X^{(i)} = (X_{\cdot\cdot 1}, X_{\cdot\cdot 2})'$ for subject i. We can view $\lambda' X_{(i)}$ as the observations and assume that it has the same variance σ^2 across the groups. The variance of $\hat{\delta}$ is thus $(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}})\sigma^2$, and σ^2 can be estimated with the pooled variance. Hence we can test if $\delta = 0$ with a t-test with the distribution

$$T = \frac{\hat{\delta}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\hat{\sigma}} \sim t_{n_{11}+n_{12}+n_{21}+n_{22}-4}\left(\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{-1/2} \theta\right),$$

where $\theta = \delta/\sigma$. We have reanalyzed the replication data and confirmed that our test is equivalent. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O = \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 6 & 6 \\ 6 & 6 \end{bmatrix}.$$

Study 49: Albarracín et al. (2008) The study has a one-way (two-cell) between-subject design. In other words, we can use a two-sample t-test on the observations. We have reanalyzed the replication data and confirmed that our test is equivalent. Hence the test statistic T follows $T \sim t_{n_1+n_2-2}\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2} \theta\right)$, where $\theta = (\mu_1 - \mu_2)/\sigma$. We have $n_{1O} + n_{2O} = 36$, $n_{1R} = 39$ and $n_{2R} = 49$, but n_{1O} and n_{2O} are unavailable.

Study 52: Centerbar et al. (2008) The original study has a 2×3 mixed design with the first factor being between-subject and the second factor being within-subject. We are interested in the interaction. The original study compared a linear trend which amounts to applying a linear contrast for the second factor, and reported a $F(1, \cdot)$-statistic. The replication code (`https://osf.io/g29pw/`) does not use a linear contrast and produces a $F(2, \cdot)$-statistic instead. However the $F(1, \cdot)$-statistic for the replication, as recorded on the RP:P dataset can be produced if we apply the contrast, which amounts to omitting the second level of the second factor. We will thus consider this, a 2×2 mixed design, as the setup. This is the same setup as Study 4 and thus we define δ, θ, σ^2 similarly. We have $n_{1O} + n_{2O} = 133$, $n_{1R} = 59$ and $n_{2R} = 54$.

Study 53: Amodio et al. (2008) The study has a 2×2 mixed design with the first factor being between-subject and the second factor being within-subject. We are interested in the interaction. This is the same setup as Study 4, and hence we defined δ, θ and σ^2 similarly. We have $n_{1O} = 16$, $n_{2O} = 15$, $n_{1R} = 29$ and $n_{2R} = 46$.

Study 56: van Dijk et al. (2008) The original study has a 2×2 between-subject design and we are interested in the main effect of the first factor on one of the second factor's level. This can be still be thought of as a two-sample t-test, except variance is estimated by pooling all four groups. The replication used a two-sample t-test that directly compare those two cells. We have $n_{11O} + n_{12O} + n_{21O} + n_{22O} = 103$ and $n_{11R} = n_{21R} = 20$, but n_{11O} and n_{21O} are not available.

Study 58: Lemay and Clark (2008) The study has a 2×2 mixed design with the first factor being within-subject and the second factor being between-subject. We are interested in the interaction. This is the same setup as Study 4, and hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 123$, $n_{2O} = 63$, $n_{1R} = 192$ and $n_{2R} = 88$.

Study 61: Ersner-Hershfield et al. (2008) The test of interest is a two-sample t-test, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 59$, $n_{2O} = 51$, $n_{1R} = 113$ and $n_{2R} = 110$.

Study 63: Correll (2008) The study has a one-way (two-cell) between-subject design, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} \approx \frac{71}{3}$, $n_{2O} \approx \frac{71\times 2}{3}$, $n_{1R} = 48$ and $n_{2R} = 100$.

Study 65: Exline et al. (2008) The study has a 2×2 between-subject design and we are interested in the interaction. In other words, we want to test if

$$\delta = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$$

is zero. A natural estimate for δ is

$$\hat{\delta} = \frac{1}{n_{11}} \sum_{i \in I_{11}} X_i - \frac{1}{n_{12}} \sum_{i \in I_{12}} X_i - \frac{1}{n_{21}} \sum_{i \in I_{21}} X_i + \frac{1}{n_{22}} \sum_{i \in I_{22}} X_i.$$

If we assume that the observations has the same variance σ^2 across the groups, then the variance of $\hat{\delta}$ is thus $(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}})\sigma^2$, and σ^2 can be estimated with the pooled variance. Hence we can test if $\delta = 0$ with a t-test using the distribution

$$T = \frac{\hat{\delta}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\hat{\sigma}} \sim t_{n_{11}+n_{12}+n_{21}+n_{22}-4}\left(\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{-1/2}\theta\right),$$

where $\theta = \delta/\sigma$. We have reanalyzed the replication data and confirmed that our test is equivalent. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O = \begin{bmatrix} 25 & 20 \\ 58 & 52 \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 31 & 33 \\ 24 & 47 \end{bmatrix}.$$

Study 68: Risen and Gilovich (2008) The study has a 2×2 between-subject design and we are interested in the interaction. This is the same setup as in Study 65, hence we define δ, θ and σ^2 similarly. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O = \begin{bmatrix} 40 & 40 \\ 40 & 40 \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 68 & 59 \\ 46 & 53 \end{bmatrix}.$$

Study 71: Stanovich and West (2008) The test of interest is a two-sample t-test, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 116$, $n_{2O} = 259$, $n_{1R} = 77$ and $n_{2R} = 100$.

Study 72: Blankenship and Wegener (2008) The study has a 2 × 2 between-subject design and we are interested in the interaction. This is the same setup as in Study 65, hence we define δ, θ and σ^2 similarly. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O \approx \begin{bmatrix} \frac{261}{4} & \frac{261}{4} \\ \frac{261}{4} & \frac{261}{4} \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 64 & 61 \\ 62 & 64 \end{bmatrix}.$$

Study 81: Shnabel and Nadler (2008) The study has a 2 × 2 between-subject design and we are interested in the interaction. This is the same setup as in Study 65, hence we define δ, θ and σ^2 similarly. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O \approx \begin{bmatrix} \frac{94}{4} & \frac{94}{4} \\ \frac{94}{4} & \frac{94}{4} \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 33 & 34 \\ 37 & 37 \end{bmatrix}.$$

Study 87: Goff et al. (2008) The study has a 2 × 2 between-subject design and we are interested in the interaction. This is the same setup as in Study 65, hence we define δ, θ and σ^2 similarly. We have

$$n_{11O} + n_{12O} + n_{21O} + n_{22O} = 55, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 12 & 14 \\ 13 & 12 \end{bmatrix},$$

but the original cell sizes are unavailable.

Study 94: McCrea (2008) The test of interest is a two-sample t-test, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 13$, $n_{2O} = 15$, $n_{1R} = 29$ and $n_{2R} = 32$.

Study 97: Purdie-Vaughns et al. (2008) The study has a 2 × 2 between-subject design and we are interested in the interaction. This is the same setup as in Study 65, and hence we define δ, θ and σ^2 similarly. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O \approx \begin{bmatrix} \frac{37}{2} & \frac{37}{2} \\ \frac{40}{2} & \frac{40}{2} \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 691 & 679 \\ 57 & 63 \end{bmatrix},$$

Study 106: Dessalegn and Landau (2008) The original study has a 2 × 3 mixed design with the first factor being between-subject and the second factor being within-subject. We are interested in the main effect of the first factor. This is equivalent to a two-sample t-test with all samples in each group pooled together. The replication used a two-sample t-test directly, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 24$, $n_{2O} = 12$, $n_{1R} = 31$ and $n_{2R} = 16$.

Study 107: Eitam et al. (2008) The test of interest is a two-sample t-test, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 41$, $n_{2O} = 45$, $n_{1R} = 79$ and $n_{2R} = 79$.

Study 110: Farris et al. (2008) While the original study used general linear model, the replication supposedly used the same model and ended up using ANOVA. We assume this is the correct analysis for the original study as well.

The replication study has a 2×2 mixed design with the first factor being within-subject and the second factor being between-subject. We are interested in the interaction. This is the same setup as Study 4, and hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 178$, $n_{2O} = 102$, $n_{1R} = 65$ and $n_{2R} = 79$.

Study 111: Janiszewski and Uy (2008) The study has a 2×2 between-subject design. However the RP:P dataset does not match the original article in that the effect being investigated is in fact the interaction. Fortunately, since the replication data is available, we can reanalyze to obtain an interaction effect of $F(1, 116) = 0.01318$, $p = 0.9088$.

This is the same setup as in Study 65, hence we define δ, θ and σ^2 similarly. We have

$$n_{11O} + n_{12O} + n_{21O} + n_{22O} = 59, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 30 & 30 \\ 30 & 30 \end{bmatrix},$$

but the original cell sizes are unavailable.

Study 113: Armor et al. (2008) The test of interest is a one-sample t-test. We are testing if the observations has mean 0. Hence the test statistic T follows $T \sim t_{n-1}(\sqrt{n}\theta)$, where $\theta = \mu/\sigma$ and σ^2 is the variance of individual observations. We have $n_O = 125$ and $n_R = 176$.

Study 114: Addis et al. (2008) The study has a $2 \times 2 \times 2$ mixed design with the first two factors being within-subject and the third factor being between-subject. We are interested in the main effect of the third factor. In other words, we want to test if

$$\delta = \mu_{**1} - \mu_{**2}$$

is zero. A natural estimate is

$$\hat{\delta} = \frac{1}{n_1} \sum_{i \in I_1} \lambda' X^{(i)} - \frac{1}{n_2} \sum_{i \in I_2} \lambda' X^{(i)}$$

where $\lambda' = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and $X^{(i)}$ is the observation $(X_{11\cdot}, X_{12\cdot}, X_{21\cdot}, X_{22\cdot})'$ for subject i. We can view $\lambda' X^{(i)}$ as the observations and perform a two-sample t-test. We have reanalyzed the replication data and confirmed that our test is equivalent.

Hence the test statistic T follows $T \sim t_{n_1+n_2-2}\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2} \theta\right)$, where $\theta = \delta/\sigma$ and σ^2 is the variance of $\lambda' X^{(i)}$. We have $n_{1O} = n_{2O} = n_{1R} = n_{2R} = 16$.

Study 115: Nurmsoo and Bloom (2008) The test of interest is a one-sample t-test, same as Study 113. We have $n_O = 32$ and $n_R = 8$.

Study 116: Vul and Pashler (2008) The test of interest is a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 173$ and $n_R = 140$ (Steegen et al., 2014).

Study 118: Masicampo and Baumeister (2008) The study has a $2 \times 2 \times 2$ between-subject design and we are interested in a particular contrast. We want to test if

$$\delta = -(\mu_{111} - \mu_{112}) - (\mu_{121} - \mu_{122}) - (\mu_{221} - \mu_{222}) + 3(\mu_{211} - \mu_{212}).$$

is zero. A natural estimate $\hat{\delta}$ is obtained by estimating each population cell mean with its sample cell mean. If we assume that the observations has the same variance σ^2 across the groups, then the variance of $\hat{\delta}$ is thus $(\frac{1}{n_{111}} + \frac{1}{n_{112}} + \frac{1}{n_{121}} + \frac{1}{n_{122}} + \frac{9}{n_{211}} + \frac{9}{n_{212}} + \frac{1}{n_{221}} + \frac{1}{n_{222}})\sigma^2$, and σ^2 can be estimated with the pooled variance. Hence we can test if $\delta = 0$ with a t-test using the distribution

$$\begin{aligned} T &= \frac{\hat{\delta}}{\sqrt{\frac{1}{n_{111}} + \frac{1}{n_{112}} + \frac{1}{n_{121}} + \frac{1}{n_{122}} + \frac{9}{n_{211}} + \frac{9}{n_{212}} + \frac{1}{n_{221}} + \frac{1}{n_{222}}}\hat{\sigma}} \\ &\sim t_{n-8}\left(\left(\frac{1}{n_{111}} + \frac{1}{n_{112}} + \frac{1}{n_{121}} + \frac{1}{n_{122}} + \frac{9}{n_{211}} + \frac{9}{n_{212}} + \frac{1}{n_{221}} + \frac{1}{n_{222}}\right)^{-1/2}\theta\right) \end{aligned}$$

where $\theta = \delta/\sigma$. We have reanalyzed the replication data and confirmed that our test is equivalent. We have

$$\begin{bmatrix} n_{111} & n_{112} & n_{121} & n_{122} \\ n_{211} & n_{212} & n_{221} & n_{222} \end{bmatrix}_O = \begin{bmatrix} 14 & 20 & 13 & 14 \\ 20 & 13 & 13 & 13 \end{bmatrix}$$
$$\begin{bmatrix} n_{111} & n_{112} & n_{121} & n_{122} \\ n_{211} & n_{212} & n_{221} & n_{222} \end{bmatrix}_R = \begin{bmatrix} 23 & 21 & 23 & 18 \\ 13 & 22 & 21 & 25 \end{bmatrix}$$

Study 121: Tabibnia et al. (2008) The original test of interest is a one-sample t-test while the replication test of interest is a z-test. We will the latter is the same test, except that the variance of the observation is known instead of estimated. This is the same setup as Study 29, and thus we define δ, θ and σ^2 similarly. We have $n_O = 12$ and $n_R = 24$.

Study 122: Alvarez and Oliva (2008) The test of interest is a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 8$ and $n_R = 17$.

Study 124: Lau et al. (2008) The study has a 2×2 mixed design with the first factor being between-subject and the second factor being within-subject. We are interested in the interaction. This is the same setup as Study 4, and hence we define δ, θ and σ^2 similarly. We have $n_{1O} + n_{2O} = 36$, $n_{1R} = 34$ and $n_{2R} = 36$, but n_{1O} and n_{2O} are unavailable.

Study 127: Winawer et al. (2008) The test of interest is a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 29$ and $n_R = 26$.

Study 133: Nairne et al. (2008) The study has a one-way (two-cell) within-subject design. Equivalently, we can use a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 24$ and $n_R = 38$.

Study 136: Vohs and Schooler (2008) The test of interest is a two-sample t-test, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} \approx n_{2O} \approx 15$ and $n_{1R} = n_{2R} = 29$.

Study 145: Halevy et al. (2008) The study has a 2×2 between-subject design and we are interested in the main effect of the first factor. In other words, we want to test if

$$\delta = \mu_{1*} - \mu_{2*}$$

is zero. A natural estimate of δ is

$$\hat{\delta} = \frac{1}{n_{11}} \sum_{i \in I_{11}} \frac{1}{2} X_i + \frac{1}{n_{12}} \sum_{i \in I_{12}} \frac{1}{2} X_i - \frac{1}{n_{21}} \sum_{i \in I_{21}} \frac{1}{2} X_i - \frac{1}{n_{22}} \sum_{i \in I_{22}} \frac{1}{2} X_i$$

where X_i are the observations. We can view $\frac{1}{2} X_i$ as the observations and assume that it has the same variance σ^2 across the groups. The variance of $\hat{\delta}$ is thus $(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}})\sigma^2)$ and σ^2 can be estimated with the pooled variance. Hence we can test if $\delta = 0$ with a t-test with the distribution

$$T = \frac{\hat{\delta}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}})\hat{\sigma}}} \sim t_{n_{11}+n_{12}+n_{21}+n_{22}-4} \left(\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{-1/2} \theta \right),$$

where $\theta = \delta/\sigma$. We have reanalyzed the replication data and confirmed that our test is equivalent. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O \approx \begin{bmatrix} 20 & 20 \\ 20 & 20 \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 10 & 10 \\ 10 & 10 \end{bmatrix}.$$

Study 146: Janssen et al. (2008) The test of interest is a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 15$, $n_R = 12$.

Study 148 and 149: Bressan and Stranieri (2008) The study has a $2 \times 2 \times 2$ mixed design with the first factor being within-subject and the second and third factors being between-subject. We are interested in the three-way interaction. In other words, we want to test if

$$\delta = \mu_{111} - \mu_{112} - \mu_{121} - \mu_{211} + \mu_{122} + \mu_{212} + \mu_{221} - \mu_{222}$$

is zero. A natural estimate of δ is

$$\hat{\delta} = \frac{1}{n_{11}} \sum_{i \in I_{11}} \lambda' X^{(i)} - \frac{1}{n_{12}} \sum_{i \in I_{12}} \lambda' X^{(i)} - \frac{1}{n_{21}} \sum_{i \in I_{21}} \lambda' X^{(i)} + \frac{1}{n_{22}} \sum_{i \in I_{22}} \lambda' X^{(i)},$$

where $\lambda' = (1, -1)$ and $X^{(i)}$ is the observation $(X_{1\cdot\cdot}, X_{2\cdot\cdot})'$ for subject i. We can view $\lambda' X^{(i)}$ as the observations and assume that it has the same variance σ^2 across the groups. The variance of $\hat{\delta}$ is thus $(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}})\sigma$ and σ^2 can be estimated with the pooled variance. Hence we can test if $\delta = 0$ with a t-test with the distribution

$$T = \frac{\hat{\delta}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\hat{\sigma}} \sim t_{n_{11}+n_{12}+n_{21}+n_{22}-4} \left(\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{-1/2} \theta \right),$$

where $\theta = \delta/\sigma$. We have reanalyzed the replication data and confirmed that our test is equivalent. For the original study, we have $n_{11O} + n_{12O} = 101$, $n_{21O} + n_{22O} = 97$ but the individual cell sizes are not known. For the two replications, we have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_{\text{Study 148}} = \begin{bmatrix} 93 & 50 \\ 75 & 45 \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_{\text{Study 149}} = \begin{bmatrix} 107 & 70 \\ 83 & 58 \end{bmatrix}.$$

Study 150: Forti and Humphreys (2008) The study has a $2 \times 2 \times 2$ within-subject design and we are interested in the two-way interaction between the second and third factors. This is the same setup as Study 1 and hence we define δ, θ and σ^2 similarly. We have $n_O = 14$ and $n_R = 19$.

Study 151: Schnall et al. (2008) The study has a one-way (two-cell) between-subject design, same as Study 49. Hence we define δ, θ and σ^2 similarly. We have $n_{1O} = 21$, $n_{2O} = 22$, $n_{1R} = 68$ and $n_{2R} = 58$.

Study 153: Palmer and Ghose (2008) The test of interest is a one-sample t-test, same as Study 113. We have $n_O = n_R = 8$.

Study 158: Goschke and Dreisbach (2008) The study has a $2 \times 2 \times 2$ mixed design with the first two factors being within-subject and the third factor being between-subject. We are interested in the two-way interaction between the first two factors. In other words, we want to test if

$$\delta = \mu_{11*} - \mu_{12*} - \mu_{21*} + \mu_{22*}$$

is zero. A natural estimate of δ is

$$\hat{\delta} = \frac{1}{n_1} \sum_{i \in I_1} \lambda' X^{(i)} + \frac{1}{n_2} \sum_{i \in I_2} \lambda' X^{(i)}$$

where $\lambda' = (\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2})$ and $X^{(i)}$ is the observation $(X_{11\cdot}, X_{12\cdot}, X_{21\cdot}, X_{22\cdot})'$ for subject i. We can view $\lambda' X^{(i)}$ as the observations and assume that it has the same variance σ^2 across the groups. The variance of $\hat{\delta}$ is thus $(\frac{1}{n_1} + \frac{1}{n_2})\sigma^2$, and σ^2 can be estimated with the pooled variance. Hence we can test if $\delta = 0$ with a t-test with the distribution

$$T = \frac{\hat{\delta}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\hat{\sigma}} \sim t_{n_1+n_2-2}\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2} \theta\right),$$

where $\theta = \delta/\sigma$. We have reanalyzed the replication data and confirmed that our test is equivalent. We have $n_{1O} = n_{2O} = 20$, $n_{1R} = 46$ and $n_{2R} = 49$.

Study 161: LoBue and DeLoache (2008) The study has a 2×2 between-subject design and we are interested in the main effect of the second factor. This is the same setup as Study 145, and hence we define δ, θ and σ^2 similarly. We have

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_O \approx \begin{bmatrix} 12 & 12 \\ 12 & 12 \end{bmatrix}, \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}_R = \begin{bmatrix} 12 & 13 \\ 11 & 12 \end{bmatrix}.$$

Study 167: Estes et al. (2008) The study has a one-way (two-cell) within-subject design. Equivalently, we can use a paired t-test, same as Study 3. We define θ similarly, and have $n_O = 18$ and $n_R = 22$.

2 Correlations and regressions

For studies with one or more continuous independent variable, the strategy used in ANOVA above no longer applies. Suppose the true (partial) correlation coefficient is r, then we can take the effect size to be $\theta = \tanh^{-1}(r)$. By Fisher transformation, the sample (partial) correlation coefficient, R, approximately follows

$$\sqrt{n-3-p}\tanh^{-1}(R) \sim N(\sqrt{n-3-p}\,\theta, 1),$$

when p independent variables are controlled for (Fisher, 1924).

Study 39: Pleskac (2008) The test contrasts correlated sample correlation coefficients as per Meng et al. (1992). The four sample correlation coefficients, r_i, and their Fisher transformed z-statistic, Z_i are combined into a z-statistic by the formula

$$Z = (Z_1 - Z_2 + Z_3 - Z_4)\sqrt{\frac{n-3}{4(1-r_x)h}},$$

where r_x is taken to the median among correlations between the independent variables, h is given by $(1 - f\overline{r^2})/(1-\overline{r^2})$, $f = (1-r_x)/2(1-\overline{r^2})$ and $\overline{r^2}$ is the average of r_i^2. $\theta = \mathbb{E}[Z_1 - Z_2 + Z_3 - Z_4]$ is the parameter whose confidence interval is given in Meng et al. (1992), and will be the choice of effect size in our analysis. In other words, if we assume r_x and h is known, then the test statistic Z approximately has distribution

$$Z \sim N\left(\sqrt{\frac{n-3}{4(1-r_x)h}}\,\theta, 1\right).$$

We have $\left(\sqrt{\frac{n-3}{4(1-r_x)h}}\right)_O = 6.926$ and $\left(\sqrt{\frac{n-3}{4(1-r_x)h}}\right)_R = 7.824$.

Study 44: Payne et al. (2008) The test statistic comes from comparing a linear model with two independent variables and the same linear model with the interaction term added. The effect is the partial correlation, between the dependent variable and the interaction, controlling for the two independent variables. We have reanalyzed the replication data and confirmed that this test is equivalent. From Fisher transformation for partial correlation (Fisher, 1924), we have

$$\sqrt{n-5}\tanh^{-1}(R) \sim N(\sqrt{n-5}\,\theta, 1),$$

and $R_O = 0.3522$, $n_O = 71$, $R_R = -0.1502$ and $n_R = 180$.

Study 48: Cox et al. (2008) The test statistic comes from determining the coefficient in a three independent variable model with full interaction. The t-statistic provided tests for the significance of the slope in the first variable when the second variable is one SD above its mean and the third variable is one SD below its mean. We can alternatively shift the second and third variables, making this desired point zero, and run the regression. The effect is thus the partial correlation of the first variable with the dependent variable, controlling for the other two variables, the three two-way interactions and the three-way interaction. We have

$$\sqrt{n-9}\tanh^{-1}(R) \sim N(\sqrt{n-9}\,\theta, 1),$$

and $R_O = -0.2255$, $n_O = 100$, $R_R = -0.05229$ and $n_R = 200$.

Study 93: Murray et al. (2008) The test statistic comes from testing the significance of the three-way interaction of two categorical and one continuous independent variables. The effect is thus the partial correlation of the three-way interaction with the dependent variable, controlling for the three independent variables and their pairwise two-way interactions. We have reanalyzed the replication data and confirmed that this test is equivalent. We have

$$\sqrt{n-9}\tanh^{-1}(R) \sim N(\sqrt{n-9}\,\theta, 1),$$

and $R_O = 0.3175$, $n_O = 91$, $R_R = -0.1351$ and $n_R = 76$.

Study 112: McKinstry et al. (2008) The original test statistic is a $F(1,\cdot)$-statistic coming from a simple regression. The effect is the correlation of the independent variable with the dependent variable. The original study provided $R_O = -0.70$ with $n_O = 11$ points in the regression. The replication gives $R_R = -0.1707$ and $n_R = 11$.

Study 120: Hajcak and Foti (2008) Suppose the true correlation coefficient is r and the joint distribution of the variables is bivariate Gaussian. We have

$$\sqrt{n-3}\tanh^{-1}(R) \sim N(\sqrt{n-3}\,\theta, 1),$$

and $n_O = 31$ and $n_R = 43$.

Study 134: Larsen and McKibban (2008) The original test statistic is a t-statistic from a regression. The effect is the partial correlation of an independent variable with the dependent variable, controlling for another independent variable. We have

$$\sqrt{n-4}\tanh^{-1}(R) \sim N(\sqrt{n-4}\,\theta, 1),$$

and $n_O = 119$ and $n_R = 238$.

Study 154: Heine et al. (2008) Since the independent variables are in fact the same sample for all eight sample correlation coefficients, we do not agree with the analysis method implemented in `https://osf.io/akv6y/` in that the resulting test statistic is not a z-score with unit variance. Nonetheless, if we assume the analysis is done correctly here, then by Fisher transformation the test statistic will follow

$$Z \sim N(\sqrt{n-3}\,\theta, 1)$$

for some parameter θ that does not vary with sample size. We will consider this θ as the effect size. We have $n_O = 70$ and $n_R = 16$.

Study 155: Moeller et al. (2008) The test statistic is a sample correlation coefficient, same as Study 120. We have $n_O = 53$ and $n_R = 72$.

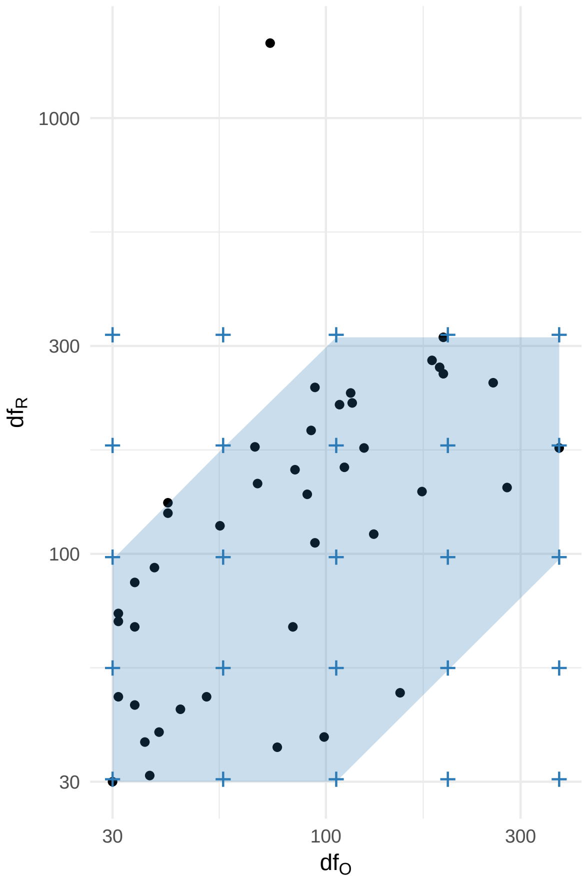

Figure 1: Degrees of freedom in the original and replication experiments where both are at least 30, on log-scale. The blue region covers all but one study pair, and hence the grid points we choose, marked as "+", are generally representative.

3 Normal approximation to t-distribution

To investigate how well the normal approximation works to t-distributions, we first consider the typical degrees of freedom for the t-distributions. Figure 1 shows the degrees of freedom in the original and replication experiments where both are at least 30. All but one study pair falls in the blue region, and hence the grid points marked by "+" are generally representative.

For each grid point, we simulate a pair of one-sided t-tests with the same effect sizes. We generate

$$T_O \sim t_{df_O}(\text{ncp}_O) 1_{\{|T_O| > t_{df_O, \alpha/2}\}} \qquad \text{and} \qquad T_R \sim t_{df_R}(\text{ncp}_R).$$

Since the original sample size is typically chosen to achieve a certain power, we assume ncp_O stays small. The type I error rate of the selective z-test is given in red in Figure 2. The type I error rate can deviate from 0.05 as the noncentrality parameter grows.

This deviation is caused by inaccuracy of approximating a noncentral t-distribution with a location-shifted standard Gaussian. We propose a finite sample correction by approximating the distribution of T_O with

$$T_O \sim t_{df_O}(\text{ncp}_O) 1_{\{|T_O| > t_{df_O, \alpha/2}\}} \approx N\left(\text{ncp}_O, 1 + \frac{2\text{ncp}_O^2}{df_O}\right) 1_{\{|T_O| > t_{df_O, \alpha/2}\}}$$

and approximating the distribution of T_R similarly. Note that the distribution of the test statistic relies on the unknown noncentrality parameter, which we replace with a plug-in estimator based on T_R: T_R can stand in for ncp_R, as well as ncp_O through the common effect size. The resulting type I error rate behaves better and is given in blue in Figure 2.

With the finite sample correction, five (11%) studies are rejected. Controlling the false discovery rate at 0.10, we apply Benjamini–Hochberg procedure (1995) and rule four (9%) replication studies as inconsistent with the original studies, namely Dodson et al. (2008); van Dijk et al. (2008); Purdie-Vaughns et al. (2008); Farris et al. (2008), generally in line with our results without the finite sample correction. Farris et al. (2008) remains rejected at familywise error rate 0.05. To check if our assumptions still hold reasonably well, we recreate Figure 2 with the effective p-value threshold of 0.004 ($= \frac{4}{46} \cdot 0.05$) used in Benjamini–Hochberg procedure and 0.001 ($= \frac{1}{46} \cdot 0.05$), given in Figure 3 and Figure 4 respectively.

For sake of completeness, we repeat the above plots specifically for the outlier (Study 97; Purdie-Vaughns et al., 2008) with exceptionally large replication degree of freedom, in Figure 5.

Our overestimate, underestimate and confidence interval for the proportion of effect sizes that declined remain the same, but we now estimate conservatively that 14 (30%) of the effect sizes declined by at least 20% with a 95% lower confidence bound of three (7%).

References

Donna Rose Addis, Alana T Wong, and Daniel L Schacter. Age-related changes in the episodic simulation of future events. *Psychological Science*, 19(1):33–41, January 2008.

Dolores Albarracín, Ian M Handley, Kenji Noguchi, Kathleen C McCulloch, Hong Li, Joshua Leeper, Rick D Brown, Allison Earl, and William P Hart. Increasing and decreasing motor

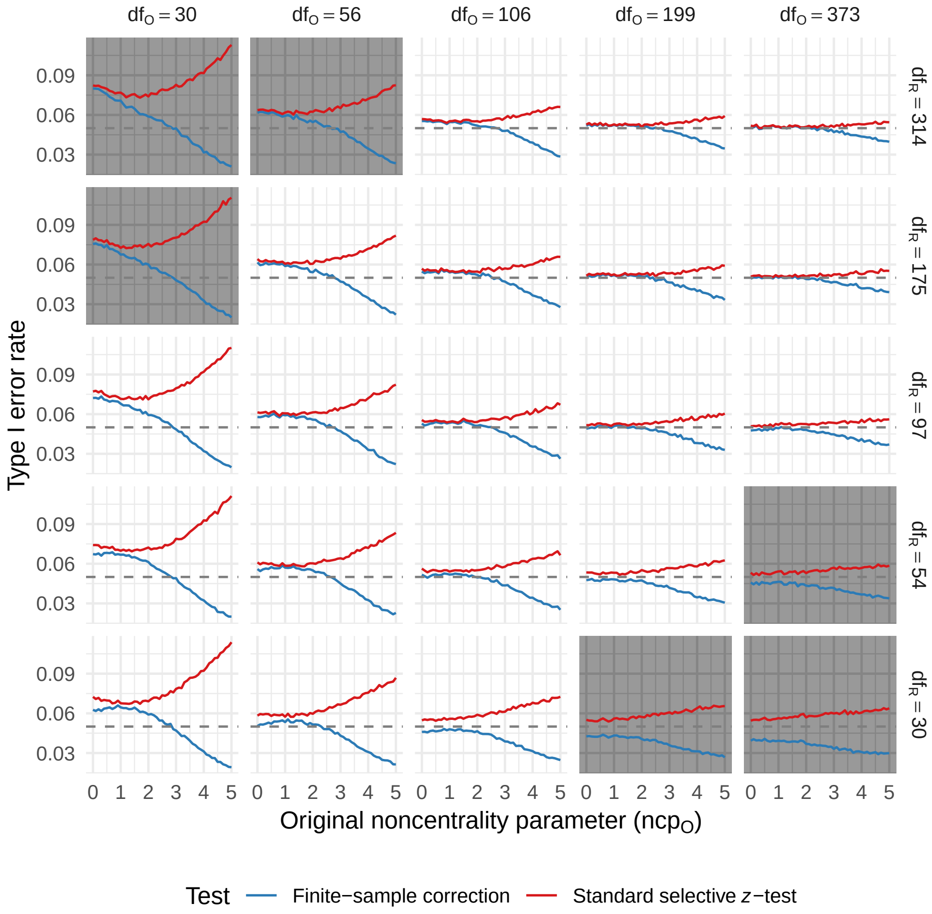

Figure 2: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red, which can deviate from 0.05 when the noncentrality parameter is large. The type I error rate of the selective z-test with our proposed finite sample correction is in blue, and stay mostly controlled. Extreme differences in degrees of freedom, as indicated by the gray background, is absent.

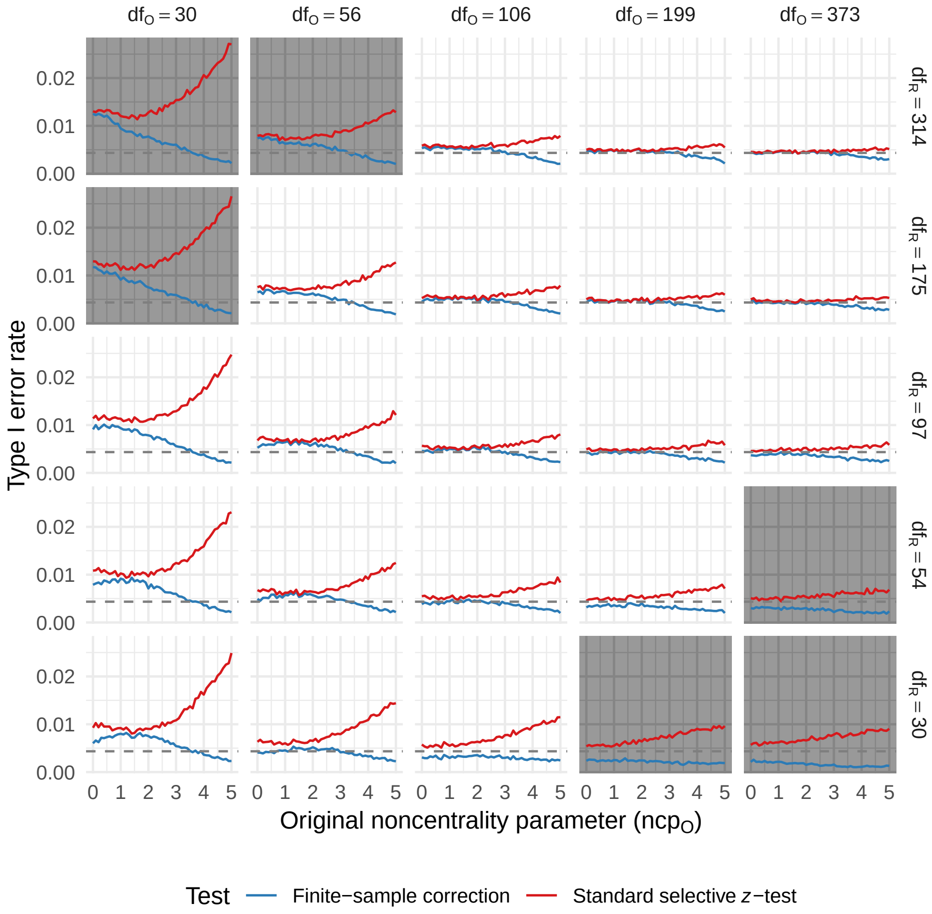

Figure 3: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red, which can deviate from 0.004 when the noncentrality parameter is large. The type I error rate of the selective z-test with our proposed finite sample correction is in blue. Extreme differences in degrees of freedom, as indicated by the gray background, is absent.

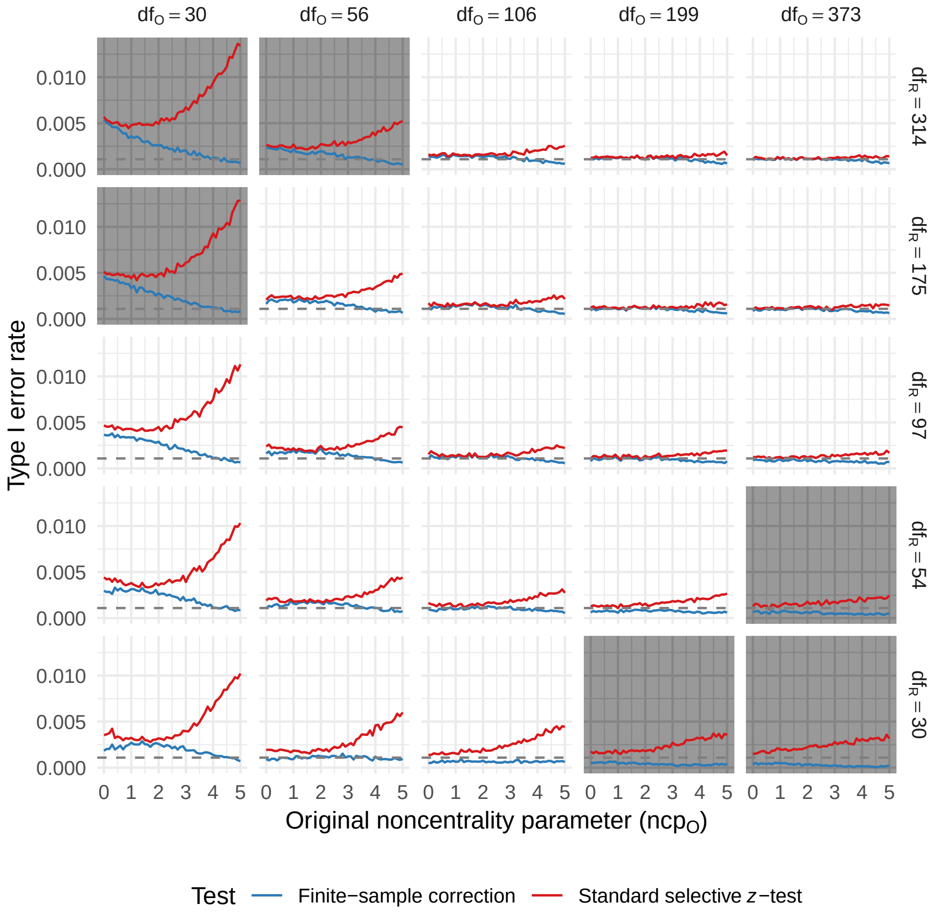

Figure 4: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red, which can deviate from 0.001 when the noncentrality parameter is large. The type I error rate of the selective z-test with our proposed finite sample correction is in blue. Extreme differences in degrees of freedom, as indicated by the gray background, is absent.

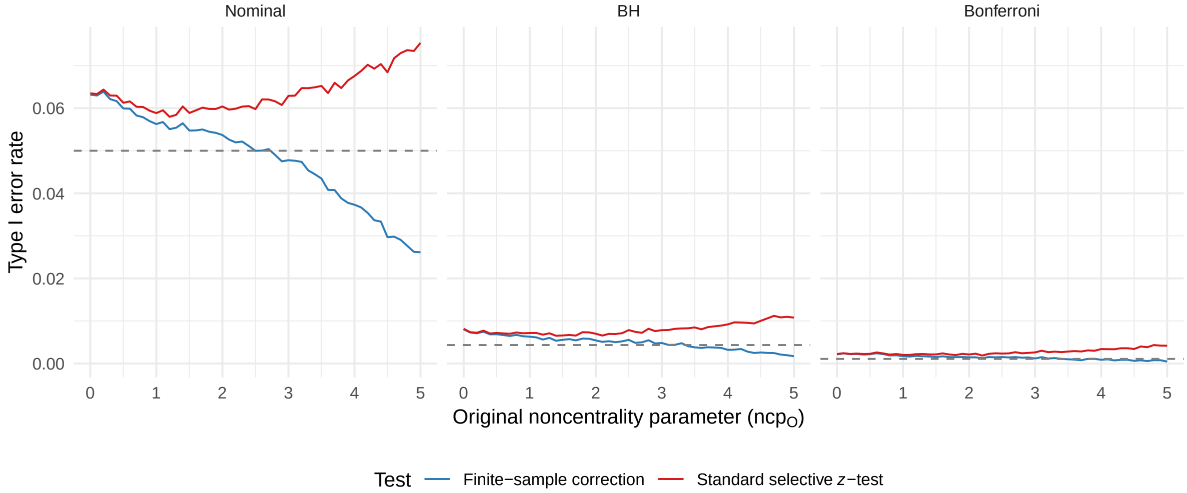

Figure 5: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red and the type I error rate of the selective z-test with our proposed finite sample correction is in blue. The error rate is evaluated for a test with the nominal level, the effective level from Benjamini–Hochberg procedure and from Bonferroni correction.

and cognitive output: A model of general action and inaction goals. *Journal of Personality and Social Psychology*, 95(3):510–523, 2008.

George A Alvarez and Aude Oliva. The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19(4):392–398, April 2008.

David M Amodio, Patricia G Devine, and Eddie Harmon-Jones. Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, 94(1):60–74, 2008.

David A Armor, Cade Massey, and Aaron M Sackett. Prescribed optimism: Is it right to be wrong about the future? *Psychological Science*, 19(4):329–331, April 2008.

Miriam Bassok, Samuel F Pedigo, and An T Oskarsson. Priming addition facts with semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2): 343–352, 2008.

C Philip Beaman, Ian Neath, and Aimée M Surprenant. Modeling distributions of immediate memory effects: No strategies needed? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):219–229, 2008.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 57(1):289–300, 1995.

Christopher J Berry, David R Shanks, and Richard N A Henson. A single-system account of the relationship between priming, recognition, and fluency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):97–111, 2008.

Kevin L Blankenship and Duane T Wegener. Opening the mind to close it: Considering a message in light of important values increases message processing and later resistance to change. *Journal of Personality and Social Psychology*, 94(2):196–213, 2008.

Paola Bressan and Debora Stranieri. The best men are (not always) already taken: Female preference for single versus attached males depends on conception risk. *Psychological Science*, 19(2):145–151, February 2008.

David B Centerbar, Gerald L Clore, Simone Schnall, and Erika Garvin. Affective incoherence: when affective concepts and embodied reactions clash. *Journal of Personality and Social Psychology*, 94(4):560–578, April 2008.

Joshua Correll. 1/f noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology*, 94(1):48–59, 2008.

Cathy R Cox, Jamie Arndt, Tom Pyszczynski, Jeff Greenberg, Abdolhossein Abdollahi, and Sheldon Solomon. Terror management and adults' attachment to their parents: The safe haven remains. *Journal of Personality and Social Psychology*, 94(4):696–717, 2008.

Banchiamlack Dessalegn and Barbara Landau. More than meets the eye: The role of language in binding and maintaining feature conjunctions. *Psychological Science*, 19(2):189–195, February 2008.

Chad S Dodson, James Darragh, and Allison Williams. Stereotypes and retrieval-provoked illusory source recollections. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):460–477, 2008.

Baruch Eitam, Ran R Hassin, and Yaacov Schul. Nonconscious goal pursuit in novel environments: The case of implicit learning. *Psychological Science*, 19(3):261–267, March 2008.

Hal Ersner-Hershfield, Joseph A Mikels, Sarah J Sullivan, and Laura L Carstensen. Poignancy: Mixed emotional experience in the face of meaningful endings. *Journal of Personality and Social Psychology*, 94(1):158–167, 2008.

Zachary Estes, Michelle Verges, and Lawrence W Barsalou. Head up, foot down: Object words orient attention to the objects' typical location. *Psychological Science*, 19(2):93–97, February 2008.

Julie Juola Exline, Roy F Baumeister, Anne L Zell, Amy J Kraft, and Charlotte V O Witvliet. Not so innocent: Does seeing one's own capability for wrongdoing predict forgiveness? *Journal of Personality and Social Psychology*, 94(3):495–515, 2008.

Simon Farrell. Multiple roles for time in short-term memory: Evidence from serial recall of order and timing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):128–145, 2008.

Coreen Farris, Teresa A Treat, Richard J Viken, and Richard M McFall. Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. *Psychological Science*, 19(4):348–354, April 2008.

Ronald Aylmer Fisher. The distribution of the partial correlation coefficient. *Metron*, 3:329–332, 1924.

Sara Forti and Glyn W Humphreys. Sensitivity to object viewpoint and action instructions during search for targets in the lower visual field. *Psychological Science*, 19(1):42–47, January 2008.

Dana Ganor-Stern and Joseph Tzelgov. Across-notation automatic numerical processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):430–437, 2008.

Phillip Atiba Goff, Claude M Steele, and Paul G Davies. The space between us: Stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology*, 94 (1):91–107, 2008.

Thomas Goschke and Gesine Dreisbach. Conflict-triggered goal shielding: Response conflicts attenuate background monitoring for prospective memory cues. *Psychological Science*, 19(1): 25–32, January 2008.

Greg Hajcak and Dan Foti. Errors are aversive: Defensive motivation and the error-related negativity. *Psychological Science*, 19(2):103–108, February 2008.

Nir Halevy, Gary Bornstein, and Lilach Sagiv. "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science*, 19(4):405–411, April 2008.

Steven J Heine, Emma E Buchtel, and Ara Norenzayan. What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science*, 19(4): 309–313, April 2008.

Chris Janiszewski and Dan Uy. Precision of the anchor influences the amount of adjustment. *Psychological Science*, 19(2):121–127, February 2008.

Niels Janssen, F-Xavier Alario, and Alfonso Caramazza. A word-order constraint on phonological activation. *Psychological Science*, 19(3):216–220, March 2008.

Jeff T Larsen and Amie R McKibban. Is Happiness Having What You Want, Wanting What You Have, or Both? *Psychological Science*, 19(4):371–377, 2008.

Grace P Lau, Aaron C Kay, and Steven J Spencer. Loving those who justify inequality: The effects of system threat on attraction to women who embody benevolent sexist ideals. *Psychological Science*, 19(1):20–21, 2008.

Edward P Lemay and Margaret S Clark. "Walking on eggshells": How expressing relationship insecurities perpetuates them. *Journal of Personality and Social Psychology*, 95(2):420–441, 2008.

Baptist Liefooghe, Pierre Barrouillet, André Vandierendonck, and Valérie Camos. Working memory costs of task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):478–494, 2008.

Vanessa LoBue and Judy S DeLoache. Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychological Science*, 19(3):284–289, March 2008.

Emer James Masicampo and Roy F Baumeister. Toward a physiology of dual-process reasoning and judgment: Lemonade, willpower, and expensive rule-based analysis. *Psychological Science*, 19(3):255–260, March 2008.

Sean M McCrea. Self-Handicapping, Excuse Making, and Counterfactual Thinking: Consequences for Self-Esteem and Future Motivation. *Journal of Personality and Social Psychology*, 95(2):274–292, July 2008.

Chris McKinstry, Rick Dale, and Michael J Spivey. Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19(1):22–24, January 2008.

Xiao-Li Meng, Robert Rosenthal, and Donald B Rubin. Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1):172–175, 1992.

Daniel Mirman and James S Magnuson. Attractor dynamics and semantic neighborhood density: processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):65–79, January 2008.

Chris Mitchell, Scott Nash, and Geoffrey Hall. The intermixed-blocked effect in human perceptual learning is not the consequence of trial spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):237–242, 2008.

Sara K Moeller, Michael D Robinson, and Darya L Zabelina. Personality dominance and preferential use of the vertical dimension of space: Evidence from spatial attention paradigms. *Psychological Science*, 19(4):355–361, April 2008.

Alison L Morris and Mary L Still. Now you see it, now you don't: Repetition blindness for nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1): 146–166, 2008.

Sandra L Murray, Jaye L Derrick, Sadie Leder, and John G Holmes. Balancing connectedness and self-protection goals in close relationships: A levels-of-processing perspective on risk regulation. *Journal of Personality and Social Psychology*, 94(3):429–459, 2008.

James S Nairne, Josefa N S Pandeirada, and Sarah R Thompson. Adaptive memory: The comparative value of survival processing. *Psychological Science*, 19(2):176–180, February 2008.

Erika Nurmsoo and Paul Bloom. Preschoolers' perspective taking in word learning: Do they blindly follow eye gaze? *Psychological Science*, 19(3):211–215, March 2008.

Klaus Oberauer. How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):439–459, 2008.

Sébastien Pacton and Pierre Perruchet. An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):80–96, 2008.

Stephen E Palmer and Tandra Ghose. Extremal edge: A powerful cue to depth perception and figure-ground organization. *Psychological Science*, 19(1):77–83, January 2008.

B Keith Payne, Melissa A Burkley, and Mark B Stokes. Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, 94 (1):16–31, 2008.

Timothy J Pleskac. Decision making and learning while taking sequential risks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):167–185, 2008.

Valerie Purdie-Vaughns, Claude M Steele, Paul G Davies, Ruth Ditlmann, and Jennifer Randall Crosby. Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, 94(4): 615–630, 2008.

Jane L Risen and Thomas Gilovich. Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, 95(2):293–307, 2008.

Ardi Roelofs. Tracing attention and the activation flow in spoken word planning using eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2): 353–368, 2008.

Lili Sahakyan, Peter F Delaney, and Emily R Waldum. Intentional forgetting is easier after two "shots" than one. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):408–414, 2008.

James R Schmidt and Derek Besner. The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):514–523, 2008.

Simone Schnall, Jennifer Benton, and Sophie Harvey. With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science*, 19(12):1219–1222, December 2008.

Nurit Shnabel and Arie Nadler. A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of Personality and Social Psychology*, 94(1):116–132, 2008.

Keith E Stanovich and Richard F West. On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94(4):672–695, 2008.

Sara Steegen, Laura Dewitte, Francis Tuerlinckx, and Wolf Vanpaemel. Measuring the crowd within again: a pre-registered replication study. *Frontiers in Psychology*, 5(786), July 2014.

Benjamin C Storm, Elizabeth Ligon Bjork, and Robert A Bjork. Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):230–236, 2008.

Golnaz Tabibnia, Ajay B Satpute, and Matthew D Lieberman. The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, 19(4):339–347, April 2008.

Nicholas B Turk-Browne, Phillip J Isola, Brian J Scholl, and Teresa A Treat. Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):399–407, 2008.

Eric van Dijk, Gerben A van Kleef, Wolfgang Steinel, and Ilja van Beest. A social functional approach to emotions in bargaining: When communicating anger pays and when it backfires. *Journal of Personality and Social Psychology*, 94(4):600–614, 2008.

Kathleen D Vohs and Jonathan W Schooler. The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, 19(1):49–54, January 2008.

Edward Vul and Harold Pashler. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647, July 2008.

Peter A White. Accounting for occurrences: A new view of the use of contingency information in causal judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):204–218, 2008.

Jonathan Winawer, Alexander C Huk, and Lera Boroditsky. A motion aftereffect from still photographs depicting motion. *Psychological Science*, 19(3):276–283, March 2008.

Melvin J Yap, David A Balota, Chi-Shing Tse, and Derek Besner. On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):495–513, 2008.