Supplementary online material: Sample Size Estimation

"Fear of death and supernatural beliefs: Developing a new Supernatural Belief Scale to test the relationship" (Jong, Bluemke, & Halberstadt, 2012)

This online appendix documents Kim's (2005) power analysis for our CFA models (Study 1).

1. Logic of statistical power to reject CFA models on the basis of null hypothesis testing

Any structural equation model, including confirmatory factor analysis (CFA), is considered acceptable as long as there is not enough statistical evidence to reject the model, as often evidenced in fit indices. When rejecting the null hypothesis (of exact, close, or adequate fit) on the basis of poor sample fit, a Type I (α) error may occur. When accepting the null hypothesis on the basis of (exact, close, or adequate) fit, a Type II (β) error may occur. "Statistical power" is the probability of rejecting a false null hypothesis and reflects unity minus the probability of accepting a Type II error (1-β). Accepting a model on the basis of the null hypothesis should not simply be due to lack of statistical power to detect model misspecification, so the decision criterion (model fit) should be strict enough. At the same time, to avoid a Type I error, the decision criterion needs to be lenient enough.

2. Specifying the degree of non-centrality for prospective power

Use of rules of thumb ($N > 200$) for SEM has been discouraged (e.g., Goffin, 2007; Iacobucci, 2010; Steiger, 2007). Even though the population covariance matrix is unknown, the minimum sample size N_{min} to attain adequate *prospective* power can be estimated separately for the fit indices *RMSEA* and *CFI*. To estimate N_{min}, one needs to specify (a) the Type I error rate (conventional $\alpha = .05$), (b) the Type II error rate (β) or targeted power (1-β), and (c) the expected population effect size, which reflects assumptions about the noncentrality, or model misspecification, to be detected if such noncentrality exists in the population.

As regards the misspecification, fundamental concerns about using fit indices have been voiced (e.g., Barrett, 2007; McDonald, 2010), and any criterion is ultimately subjective and arbitrary, so alternative degrees of misfit or different null hypotheses might be specified. "Steiger (1989) and Browne and Cudeck (1993) suggest guidelines for the interpretation of *RMSEA*: values in the range of 0.00 to 0.05 indicate close fit, those between 0.05 and 0.08 indicate fair fit, and those between 0.08 and 0.10 indicate mediocre fit. *RMSEA* values above 0.10 indicate unacceptable fit" (MacCallum, Widaman, Preacher, & Hong, 2001, p.621). A recent trend favors slightly stricter criteria (Hu & Bentler, 1999; Steiger, 2007). None of the cutoffs take into account that target value of *RMSEA* = .00 may still not be strict enough to arrive at a 90% rejection rate, while values *RMSEA* > .08 might be empirically obtained despite the model adequately describing the covariance matrix in the population (Chen, Curran, Bollen, Kirby, & Paxton, 2008). We deemed that an acceptable model would have to satisfy at least minimum criteria, so we targeted at power to detect a population *RMSEA* of .08 and *CFI* of .95. Stricter criteria might be required to discriminate between models that actually reach those values. Note that despite their numerical resemblance on the surface, *RMSEA* of .05 does not represent the same degree of misfit as *CFI* of .95 (Kim, 2005).

As regards the statistical power, Cohen (1988) suggested aiming at .80, yet when weighing Type II errors somewhat more importantly in SEM/CFA than in common null hypothesis significance testing, power of .90 might be justified as well (as one reviewer suggested). N_{min} was estimated for two levels of prospective power (.80 and .90; see Table below). The logic and details on how to compute N_{min} are documented in Kim (2005).

3. Estimated minimum sample size

To determine N_{min} with any specified prospective power for *RMSEA*, no model assumptions are required to determine the degree of noncentrality from perfect fit. Then,

$$N_{\min} = \frac{\delta_{1-\beta}}{\varepsilon^2 df} + 1,$$

with ε = .08 and $\delta_{1-\beta}$ according to the critical, *df*-specific noncentrality parameter (Kim, 2005).

To determine $N_{\min}$ with any specified prospective power for *CFI*, additional model assumptions about the expected population covariance matrix are required to determine the degree of noncentrality from an independence model of uncorrelated variables. We assumed model-specific factor loadings and factor matrices that were inspired by the EFA results (cf. Pretest). For M1 (unidimensional model), we assumed factor loadings of λ = .80 on average for the main factor (slightly more conservative than the empirical average of .825 in EFA). For M4 (orthogonal method plus content factors) we assumed λ_1 = .80, yet additionally λ_2 = .40 for the loadings on the method factor and λ_3 = .30 for loadings on the content facets. For M8 (correlated content factors), we assumed λ = .80 for each of the five factors and factor intercorrelations of Φ = .80 to reflect the strong first factor. Then,

$$N_{\min} = \frac{\delta_{1-\beta} + df_B(1 - CFI)}{F_B(1 - CFI)} + 1,$$

with *CFI* = .95, df_B = degrees of freedom of the baseline model (45 for 10 variables), F_B = −log(det[ρ]) with ρ = correlation matrix (as implied by factor loadings and factor inter-correlations), and $\delta_{1-\beta}$ according to the *df*-specific critical noncentrality parameter. The table shows $N_{\min}$ as a function of power, misfit, and model type.

Model	***df***	**(1-β)**	N_{min} **(*RMSEA* > .08)**	N_{min} **(*CFI* > .95)**
M1	35	0.80	118	79
		0.90	145	96
M4	27	0.80	138	61
		0.90	170	75
M8	25	0.80	144	109
		0.90	178	133

6. Evaluation of statistical power and acceptance of model

The fit indices for the models in Study 1 clearly indicated that the null hypothesis had to be rejected regularly ($N = 213$), because all of the alternative models were far from adequately reproducing the empirical covariance matrix (cf. Table 3). Even in the case of the finally accepted model (M4/M5), there remains a non-negligible amount of unexplained variance, as evident in the significant, yet sample-size sensitive χ^2-statistic. Evidently, power to detect misfit was not a problem in our analysis, so sample size was not too small, and the study was not underpowered. In conclusion, due to the observed misfit, none of the rejected models is at risk of being falsely accepted due to a Type II error and subsequently qualified as displaying "exact", "close", or "adequate" fit. If anything, Type I errors might have occurred when rejecting these models. Sample size was not only sufficient to reject these models, but the *observed power* for the *observed misfit* approximated unity when checking the empirical *RMSEA*s and *CFI*s with Kim's formulas.

The only model that might run the risk of a Type II error is the accepted model M4/M5, yet its acceptance is not based on strict accept-fail logic (Bentler, 2007), but on comparative testing of alternative models on the same data (Lance, Butts, & Michels, 2006). Compared to the alternative models, M4/M5 reproduces the covariance matrix best (lowest χ^2, lowest χ^2/df-ratio, lowest *SRMR*), it is the most accurate, while at the same time parsimonious model (lowest *AIC*, *TLI* > .95), and it is the only model for which the highest *CFI* (> .95) and the lowest *RMSEA* (< .08) were observed. *RMSEA* was larger than what some researchers might expect from a closely fitting model when adhering religiously to arbitrary cutoffs, yet it might have been so due to its property to be conservatively high at relatively low numbers of participants (Fan & Sivo, 2005; Kenny & McCoach, 2003; Hu & Bentler, 1998; Chen et al., 2008). *RMSEA* improved considerably when the model was tested on the combined sample (Study 1 and 2, $N = 360$), *RMSEA* = .064 (p-close = .064). Other fit indices improved likewise, *CFA* = .978, *TLI* = .963, *SRMR* = .029.

References

Barrett, P. (2007). Structural equation modelling: Adjuging model fit. *Personality and Individual Differences, 42,* 815-824.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42,* 825-829.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In: K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research, 36,* 462-494.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components. *Structural Equation Modeling,12,* 343-367.

Goffin, R. D. (2007). Assessing the adequacy of structural equation models: Golden rules and editorial policies. *Personality and Individual Differences,42,* 831-839.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424-453.

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

Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology, 20,* 90-98.

Kenny, D. A., & McCoach, B. D. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling,10,* 333-351.

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling, 12,* 368-390.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9,* 202-220.

MacCallum, R. C., Widaman, K. F., Preacher, K., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36,* 611-637.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science, 5,* 675-686.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173-180.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42,* 893-898.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors.* Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.