Supplemental material for:
"The Frequency of Excess Success for Articles in *Psychological Science*"

Gregory Francis
Purdue University
gfrancis@purdue.edu

Estimating the Probability of Experimental Success

A key part of the TES analysis involves estimating the probability of success for hypothesis tests. For some cases calculating the success probability (power) is straightforward and easily computed with software packages (Champely, 2009; Lenth, 2009). More commonly it is difficult or impossible to estimate the joint success probability of an experiment with the reported statistics for several measures, and the best that can be generated is a smallest upper limit. For example, research studies often investigate more than one dependent variable with a single sample and base a theoretical conclusion on a significant effect (or lack thereof) for each dependent variable. It is a mathematical necessity that the joint probability of success for all dependent variables producing a desired result must be lower than the probability of any one variable producing a desired result. The joint success probability can sometimes be estimated if the data summary includes the correlations between dependent variables, but such information is rarely published. An upper limit estimate of the joint success probability for multiple dependent variables can be established by the smallest individual success probability of the reported effects. Thus, if three hypothesis tests from a single sample all reject the null and indicate power values of .9, .7, and .6, then the tightest upper limit of the joint power of all three tests is .6, which is the smallest individual power.

In other cases, a single ANOVA or set of *t*-tests may impose multiple restrictions on the statistical results (e.g., a significant interaction and some, but not all, contrasts or main effects). The probability of a similar experiment reproducing a required pattern of statistical results must consider the multiple relationships between those effects. As a general principle, the more requirements placed on the outcome of an experiment, the lower the probability of the experiment to produce that pattern. Importantly, these ideas slightly generalize the concept of power to include some desired non-significant findings. What matters most for the TES analysis is the probability of the desired pattern of significant and non-significant results. Whether the probability of the desired pattern can be estimated depends on the design of the experiment and the statistics that are reported.

For between-subject designs, it is often possible to use the sample means and standard deviations (these values are often provided in the text or can be extracted from figures) to generate simulated experiments that draw new random samples and then recreate the hypothesis tests on the simulated experiments (R Development Core Team, 2011). Such simulations can then estimate the proportion of experiments that produce the desired pattern of significant and non-significant results. The proportion is then an estimate of the joint success probability.

Reports of within-subjects analyses sometimes lack sufficient summary statistics to estimate power. For a dependent design with two measures, power is often easily computed

from a given *t* or *F* value, but with more than two measures it is necessary to know the effect size (or means and standard deviations) and the correlations between the measurements, and such correlations are typically not reported. Depending on the details of the analyses, it is sometimes possible to estimate an upper limit of power, but in other cases it is not possible to provide any estimate for a within-subjects design. This difficulty is a common reason for excluding articles from the TES analysis.

For all of the estimated success probabilities, it was assumed that the original hypothesis tests were appropriate for the reported data (e.g., the data were randomly sampled from normal distributions with a common variance). When sample sizes were not given for independent samples, the TES analysis used a nearly balanced design, which tends to maximize power.

The TES analysis is agnostic about whether the experimental variables truly measure the effects of interest or whether the manipulations were effective and appropriate for the investigation. In a sense, the TES analysis is content-free in that it only critiques the statistical aspects of a set of experiments. It would be perfectly reasonable for an experiment set that satisfies the TES analysis to be further critiqued on grounds of validity and generality. On the other hand, an experiment set should engender skepticism if it does not satisfy the TES analysis, regardless of measurement validity and generality.

Details of the TES analysis for each article in Table 2 of the main article can be found below. The files accompanying this document include an individual directory for each article (ordered by year) that includes all calculations of the effect sizes (Excel files) and any R scripts that were used to estimate power. Each directory includes a text file that identifies the selected statistics and provides a few comments.

Articles Not Analysed

Of the 951 articles published in *Psychological Science* during 2009-2012, there were 79 articles that had four or more experiments. Each of these 79 articles was examined to see if the TES analysis could be performed with the information provided in an article and its supplemental material. Articles were excluded from the TES analysis if it was not possible to compute the success probability for at least four experiments. Specific reasons that success probability could not be calculated and thereby for exclusion from the TES analysis included the following, and some articles were excluded for multiple reasons.

- The experiments did not analyze the data with hypothesis testing. [8 articles]
- The judgment of statistical significance was based on values of p_{rep} (Killeen, 2005). There appears to be no established criterion for p_{rep}, so it is not possible to estimate power for these experiments. [9 articles]
- The data were analyzed with a Bayesian method. [1 article]
- The article was unclear about which statistical tests were related to the theoretical conclusions. [3 articles]
- The reported statistical analyses lacked sufficient detail to estimate power (e.g., a complex ANOVA that did not report an effect size or correlations for a within-subjects design). [19 articles]

Table S1 lists each of the 35 excluded experiments and the reasons for their exclusion from the TES analysis. Further details are available upon request.

Table S1: These articles included four or more experiments but could not be fully analysed with the TES because it was not possible to estimate success probability for at least four experiments.

Year	Authors	Reason not being fully analysed by TES
2012	Chen & Hegdé	Several experiments list only $p<.05$, which is not enough information to compute power.
2012	Endress & Potter	Key hypothesis tests are across experiments, which leads to complicated relationships between tests. Including the supplemental material, there are more than 130 hypothesis tests, and it is often not clear how they relate to the theoretical ideas.
2012	Jarvstad, Rushton, Warren & Hahn	They key results for the theory were not analysed with hypothesis tests.
2012	Sussman & Shafir	Many hypothesis tests use a Wilcoxon test. Given the tests' nonparametric properties, it is not clear how to estimate power.
2012	Waytz & Young	Experiment 3 uses *t*-test contrasts based on a within-subjects ANOVA. It is not clear how to compute power for a situation like this without more information. Experiment 4 is similar.
2011	Arneson, Sackett & Beatty	Not really a use of hypothesis tests. The authors fit models to four large data sets and looked at differences between model fits.
2011	Beneson, Markovits, Emery Thompson & Wrangham	No statistics provided for Study 4, which leaves only three experiments where power can be estimated.
2011	Jonikaitis & Deubel	Experiment 4 does not report details of any hypothesis tests, which leaves only three experiments.
2011	Noguchi, Handley & Albarracín	Study 3 reports an ANCOVA, but does not give the correlation between variables or adjusted means. Cannot compute power. This leaves only three experiments where power can be estimated.
2011	Rule, Rosen, Slepian & Ambady	For Study 2, *F* values (and other statistics) are not given, so power cannot be estimated. This leaves only three experiments where power can be estimated.
2011	Ungemach, Stewart & Reimers	Study 1a is based on a Wald test, for which there appears to be no way to compute *post hoc* power. Study 2 uses a complex regression and it is not clear how to estimate power from the given statistics. This leaves only two experiments where power can be estimated.

2011	Zhang & Luck	Main results are based on a Bayesian analysis (e.g., to find evidence for the null).
2010	Arnold, Erskine, Roseboom & Wallis	Most results are not analyzed with hypothesis tests.
2010	Bahrami, Vetter, Spolaore, Pagano, Butterworth & Rees	Hypothesis tests are based on within-subject ANOVAs having more than two conditions; there is no way to calculate power with the provided statistics.
2010	Chambers	Hypothesis tests are based on p_{rep}.
2010	Eyal & Epley	Hypothesis tests are based on p_{rep}.
2010	Gajewski, Philbeck, Pothier & Chichka	Experiment 2 does not include any statistics, so power cannot be calculated. This leaves only three experiments with a power estimate.
2010	Gino, Norton and Ariely	Hypothesis tests are based on p_{rep}.
2010	Jiga-Boy, Clark & Semin	Hypothesis tests are based on p_{rep}.
2010	Job, Dweck & Walton	It is not clear how to compute power for the regression in Study 4. This leaves only three tests with a power estimate.
2010	Langer, Djikic, Pirson, Madenci & Donohue	The reported analyses contain so many errors that it is not possible to compute power. Significance is sometimes claimed with $p>.05$ and reported χ^2 values do not match the reported proportions. Study 3 contains multiple *t*-tests that appear to be unabashedly selected from a broad set of possible comparisons.
2010	Linkenauger, Ramenzoni & Proffitt	Hypothesis tests are based on p_{rep}.
2010	Looser & Wheatley	The repeated measures ANOVA with more than two measures in Experiments 1b and 2 requires knowledge of the correlation between measures, which is not given. Power can only be estimated from three experiments.
2010	Lupyan, Thompson-Schill & Swingley	The hypothesis tests in Experiment 3 seem to describe the data rather than test the theoretical ideas. Some hypothesis tests compare data across experiments, and it is not clear how to estimate power both within and across experiments.
2010	McGraw, Larsen, Kahneman & Schkade	p_{rep} is used for Study 1a and 3. Studies 1b and 2 do not use hypothesis tests.
2010	Pratt, Radulescu, Guo & Abrams	Experiment 2a compares new findings to those in Experiment 1a. Because of multiple comparisons, it is difficult to estimate power. Experiments 3a and 3b use repeated measures ANOVAs that do not give enough information to compute power.

2010	Raffray & Pickering	Complicated regression for hypothesis tests makes computing power seem impossible. Hypothesis tests are for model fits.
2010	Swann, Gómez, Dovidio, Hart & Jetten	Complicated multinomial regression for hypothesis tests makes computing power seem impossible.
2010	Vul & Riche	Data analysis does not use hypothesis tests.
2009	Chen & Treisman	Hypothesis tests are based on p_{rep}.
2009	Finkbeiner & Palermo	Hypothesis tests are based on p_{rep}.
2009	Ginges, Hansen & Norenzayan	Not sure how to compute power for two of four hypothesis tests (not enough information given and use Wald test). Some hypothesis tests seem to be used as a way of describing the results, but do not contribute to an overall conclusion.
2009	Gollwitzer, Sheeran, Michalski & Seifert	For Study 2 and 3 the provided means and standard deviations do not match the reported test statistics.
2009	Hong & Shevell	Data analysis does not use hypothesis tests.
2009	Logan & Crump	Hypothesis tests are based on p_{rep}.

TES Analyses

1. Analysis of Anderson, Kraus, Galinsky & Keltner (2012) "Sociometric Status and Subjective Well-Being"

Anderson *et al.* (2012) reported four experiments that purported to show that sociometric status has a stronger impact on subjective well-being than socioeconomic status. Table S2 summarizes the statistics that contributed to the bias analysis.

Study 1 had numerous comparisons that were all consistent with the theoretical ideas. Since the statistics contributing to the multiple tests are connected, it is not possible to compute the joint power of all the tests. A tightest upper limit was identified by computing the power for an analysis reported in the supplemental material of Anderson *et al.* (2012). This test was included to rule out common method variance as an explanation of other results. It reported a significant result for a subset of the measured variables. The statistic in Table S2 is a *t* value.

Study 2 broadened the sample (from college-students to a national online sample) and measured additional variables (sociometric status, socioeconomic status, gender, ethnicity, and extraversion). The key reported comparison was between restricted and unrestricted models, and the power of that comparison is very high. As before there were several additional hypothesis tests of varying importance for the theory, but it is not possible to combine those tests with the power analysis.

Study 3 introduced an experimental manipulation of status. The key comparison was the interaction between level (high or low) of status and type (sociometric or socioeconomic) status. The interaction is significant, but it is from an experiment that is not particularly powerful; by itself an experiment like this would be expected to reject the null

about 57% of the time. Follow-up *t*-tests found a significant effect of level within the sociometric-status condition but not within the socioeconomic-status condition. The mean values of each (between subjects) group can be extracted from Figure 1 in Anderson *et al.* (2012). Algebra can be used to solve for the standard error given the mean values and the reported *t* statistics. Using the reported degrees of freedom (although they seem to be off by one) for the *t* tests, it is reasonable to assume that the sample sizes are 59 for each of the low and high level sociometric groups and 55 for each of the low and high socioeconomic groups, so the standard deviation can be estimated to be 0.753. A satisfactory result for Study 3 requires both a significant interaction and a significant difference between the low and high sociometric groups (the study also reported a significant effect of level and a non-significant difference between the low and high socioeconomic groups, but these outcomes are not central to the theory). The extra constraint on the data slightly reduces the probability of a successful experimental outcome. One hundred thousand simulated experiments were used to calculate the proportion of experiments that that rejected both the null hypothesis for the interaction and the null hypothesis for the low and high sociometric groups. As noted in Table S2, an experiment like this is expected to find significant results for both tests about 55% of the time.

Study 4 was a longitudinal study that asked students about social well-being before and after graduation. The key comparison was whether changes in social well-being corresponded with changes in sociometric status, and a regression showed a significant result, albeit with not especially high statistical power. The statistic in Table S2 is a *t* value. Another analysis compared restricted and non-restricted models for data after graduation and found a significant advantage for the non-restricted model. It is not possible to estimate the power of both results, but the model comparison effect was quite large, so probably a significant finding would almost always be generated. Thus, the power estimate in Table S2, which is based on just the regression, is probably pretty close to the estimate for both effects.

Table S2: Statistical properties of the Anderson *et al.* (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	80	2.00	.049	0.221	.499
Study 2	315	14.13	<.001	0.211	.962
Study 3		Multiple			.551
Study 4	116	2.33	.021	0.215	.632
P_{TES}					.167

Anderson *et al.* (2012) reported that each of the studies produced the pattern of results that support their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES}= .167. Since this value is greater than the .1 criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

2. Analysis of Bauer, Woilkie, Kim & Bodenhausen (2012) "Cuing Consumerism"

Bauer *et al.* (2012) reported five experiments that purported to show that situational cuing can trigger materialistic mind-sets. Table S3 summarizes the statistics that contributed to the TES analysis.

A pilot study was used to verify that materialistic aspiration increased with exposure to desirable consumer items; and the other experiments used this effect to produce differing experimental groups. The statistic in Table S3 is a *t* value.

The key comparison in Experiment 1 was between measures of depressed affect (derived from questionnaires) for participants in a consumer-cue condition versus participants in a control condition. The reported *t* value indicates a significant difference in explained error variance between the models (that the *p* value is bigger than .05 might be due to rounding in the reported statistic). Experiment 1 also included hypothesis tests for measures of anxious affect, dissatisfaction with self, and hours allocated to social activities. Some of these results seem important for the theoretical position espoused by Bauer *et al.* (2012), but it is not possible to compute the probability of all of these tests because there are unknown correlations between the measures. The power listed in Table S3 for Experiment 1, which corresponds only to the estimated power for depressed affect, is an upper limit of the joint power for all of the tests.

Experiment 2 measured similar effects using the Evaluative Movement Assessment with participants in either a "consumer" or "citizen" condition. As Table S3 shows, the *t* value was significant, but the power is not particularly high.

Experiment 3 used a scrambled-sentences task to prime consumer or neutral mind-sets. Success in this experiment required that the priming manipulation induced a difference in competitiveness across participants in the two groups. As Table S3 shows, there was a significant difference but the power is close to one half. In addition, this experiment also measured interest in pursuing social activities as a function of the priming conditions. The authors predicted, and found, a significant result for high-investment activities. This result by itself would have power close to one half, but it is not possible to compute the probability of the joint outcome of rejecting both hypotheses. A third hypothesis also showed a significant result for low-investment activities, but it is not clear that this result is related to the theoretical views of the authors. The power listed in Table S3 is an upper limit on the joint power of the effects that matter for Bauer *et al.* (2012).

Experiment 4 explored how a sense of personal responsibility was related to participants in a "consumer" or an "individual" mind-set. The main result was that participants in the consumer group expressed less personal responsibility for dealing with a described crisis. As Table S3 shows, the *t* statistic indicates a significant result, but the power is quite modest. Experiment 4 also measured several other variables and some of them seem important for the theoretical view espoused by the authors. There is no way to measure the power of the joint outcomes, so the power listed in Table S3 is an upper limit. Some of the measures that did not reject the null hypothesis were also described as being consistent with the theory, and including such non-significant findings would only further lower the power of Experiment 4.

Table S3: Statistical properties of the Bauer *et al.* (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Pilot	24, 23	2.70	.010	0.775	.738
Exp. 1	25, 25	2.00	.051	0.557	.488
Exp. 2	29, 29	2.33	.023	0.604	.618
Exp. 3	33, 33	2.04	.045	0.496	.510
Exp. 4	39, 39	2.12	.037	0.475	.545
P_{TES}					.062

Bauer *et al.* (2012) reported that each of the five studies produced a pattern of results that support their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting five studies like these to produce the desired pattern is the product of the probabilities: $P_{TES} = .062$. Since this value is less than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

3. Analysis of Birtel and Crisp (2012) "Treating Prejudice"

Birtel and Crips (2012) reported four experiments purporting to show that imagining a negative encounter with a stigmatized group leads to subsequent positive perceptions of that group. Table S4 summarizes the statistics that contributed to the TES analysis.

All of the experiments used a two-sample design. The key measure in Experiment 1 was an anxiety rating for participants imagining a positive meeting with a schizophrenic. The key comparison was a difference in this rating for participants who had previously imagined a positive or a negative interaction. Birtel and Crisp reported a significant difference, but the reported *p* value (.046) does not agree with the given sample sizes and *F* value. Given the reported properties of the experiment and data, the correct value should be slightly above the typical criterion for statistical significance, but if the reported *F* value was rounded down, then the result could be statistically significant (p=.0499). The power calculation used the *F* value reported in the paper.

Experiments 2a and 3 had a similar design but included two measures (anxiety/feelings and reported intentions for future contact with individuals in the stigmatized groups). Both measures were statistically significant. To compute the power of both measures being significant one needs to know the correlation between the variables, which was not provided in the text. The smallest upper limit is the smaller of the two power values. Experiments 2a and 3 also included a mediation analysis to show that the effect of the first imaginary exposure was mediated by anxiety/feelings. The power of the mediation can be computed, but it is not clear how to compute the joint power of the two significant results and the mediation. For the present analysis, the mediation effects were not included.

Experiment 2b was similar to Experiment 2a, except that the imagined meeting was with an "in-group" person. Imagining a negative encounter produced higher anxiety than imagining a positive encounter at Time 1, but the difference went away at Time 2 (as predicted by the theory). It is not possible to estimate the probability of the findings at both

times, so an upper limit of the joint probability is the power at Time 1. The statistic in Table S4 is an *F* value. The reported *p* value in Birtel and Crisp (2012) is .048, but it should actually be .052, which is just above statistical significance. The power calculation used a criterion of .052.

Birtel and Crisp (2012) reported that each of the studies produced the pattern of results that support their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .133. Since this value is greater than the .1 criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

Table S4: Statistical properties of the Birtel and Crisp (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	14, 15	4.21	.050	0.741	.485
Exp. 2a	16, 16	2.48	.019	0.854	.648
Exp. 2b	15, 15	4.12	.052	0.721	.486
Exp. 3	11, 11	3.36	.003	1.38	.867
P_{TES}					.133

4. Analysis of Converse and Fischbach (2012) "Instrumentality Boosts Appreciation"

Converse and Fischbach (2012) reported four experiments purporting to show that beneficiaries of help have feelings of appreciation that depend on how they perceive the helpers to be facilitating active goals. Table S5 summarizes the statistics that contributed to the TES analysis. All of the tests statistics are *t* values.

Experiment 1 compared appreciativeness ratings from participants playing a trivia game where participants could get assistance from helpers. The appreciation ratings were higher when measured while the helper was working on a question than when measured after receiving the helper's feedback.

Experiment 2 utilized a mixed design (and a different task) that also measured the views of helpers (in a between-subjects way). Beneficiaries and helpers showed opposite tendencies, with helpers expecting the beneficiaries to feel less indebted during the task than after it. Given the between subjects nature of the design, separate hypothesis tests were generated for the beneficiaries (2a) and for the helpers (2b).

Experiment 3 looked for natural variation among students who work with tutors to achieve academic goals. A feeling of appreciation was measured before and after exams. The experiment included many different additional measures that were correlated with each other in complex ways. It is difficult to identify power for the complex analyses used to make inferences from this experiment. Table S6 reports the power for only the appreciation ratings. A supplemental experiment asked subjects to describe either past or future help and then to rate their appreciation of the help. Ratings were higher for future help.

Table S5: Statistical properties of the Converse and Fishbach (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	18, 19	3.1	.004	0.998	.839
Exp. 2a	20	2.65	.016	0.569	.675
Exp. 2b	20	2.44	.025	0.524	.604
Exp. 3	21	2.43	.025	0.510	.604
Supplemental	57, 57	2.07	.041	0.385	.531
P_{TES}					.110

Converse and Fischbach (2012) reported that each of the studies produced the pattern of results that support their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .110$. Since this value is greater than the .1 criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

5. Analysis of Converse, Risen and Carter (2012) "Karmic Investment"

Converse *et al.* (2012) reported six experiments that purported to show that people were more likely to help others when placed in a situation with limited control. Table S5 summarizes the statistics that contributed to the TES analysis.

Experiment 1a compared volunteering rates for participants who wrote an essay about either a situation where people want something or about a routine situation. The test statistic listed in Table S6 is a χ^2 value, and the listed effect size is Cohen's *h* (n_1=49, p_1=0.939, n_2=46, p_2=0.783). Experiment 1b had a similar comparison (n_1=21, p_1=0.857, n_2=27, p_2=0.593). It also included an expected non-significant condition that framed the purported volunteer task as entertainment. The non-significant finding was not part of the TES analysis, and it would make little difference (the probability of finding a non-significant result would be around .95.)

Experiment 2 measured possible charity donations as a function of essay writing conditions. The key comparison was a contrast between the "wanting" essay prompt and two control conditions. The analysis included control of years of education, so the reported test statistics is an *F* value from an ANCOVA. The effect size is given as Hedges' *g*. It is not entirely clear how to compute the effect size for a contrast. The analysis converted the reported *F* value to *g* by treating the data as from a two-sample test. This approach might give a slight misrepresentation of the effect size because the degrees of freedom are slightly different for a contrast. Despite this difference, the *post hoc* power calculation appears to be valid. The estimated power that is based on the effect size was compared with simulation experiments using samples drawing from populations having the reported means and standard deviations. These simulated experiments rejected the null hypothesis with a rate that reflected the estimated power to the third decimal place. Experiment 3 had a similar

measure as Experiment 2, but only one control and different types of surveys established the different conditions.

Experiment 4 measured optimism after subjects were exposed to conditions that varied control or the beneficiary of a donation (self or others). The statistical comparison was a contrast between participants in the outside control and charity benefit against the other three conditions.

In the conclusions section, Converse *et al.* (2012) reported a "follow up" experiment that used the essay writing tasks and then measured how often participants reported karma-related words in a word completion task. The test statistic in Table S6 is a χ^2 value, and the listed effect size is Cohen's *h* (n_1=162, p_1=0.630, n_2=145 p_2=0.497).

Table S6: Statistical properties of the Converse *et al.* (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1a	49, 49	4.90	.027	0.470	.629
Exp. 1b	21, 27	4.00	.045	0.609	.553
Exp. 2	32, 62	4.07	.047	0.436	.508
Exp. 3	38, 39	4.87	.030	0.498	.578
Exp. 4	73, 219	5.42	.021	0.314	.638
Follow up	162, 145	5.58	.020	0.269	.653
P_{TES}					.043

Converse *et al.* (2012) reported that each of the studies supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all reject the null is the product of the probabilities: P_{TES} = .043. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

6. Analysis of Keysar, Hayakawa and An (2012) "Foreign-Language Effect"

Keysar *et al.* (2012) reported six experiments that purported to show that using a foreign language reduces decision-making biases such as the framing effect. Table S7 summarizes the statistics that contributed to the TES analysis. Experiment 1a had participants answer the "Asian disease" problem (framed with gains or losses) in either their native language or a (known) second language. Experiments 1b-c were replications with subjects from different languages. In each experiment participants working in their native tongue showed a framing effect (differences in choices across the frames), while participants working in their second language did not. These were independent samples, so Table S7 shows both the probability of a significant effect for the native language participants and the probability of a (theory consistent) non-significant effect for the second language participants. The test statistic is a χ^2 value and the effect size is Cohen's *h*. For the theory-consistent null results, if the null hypothesis is true, then the probability of getting a non-significant result is the complement of the Type I error rate.

Experiment 1d was a control to insure that the previously reported effects were not because participants did not carefully read the problems when using a second language. A third frame type was added with a highly preferable choice, and it was very frequently selected even though all subjects read it in a second language. Power for this experiment was estimated with simulated experiments that drew samples from populations having the reported proportions. An additional hypothesis test noted the expected non-significant difference between the gain and loss frames.

Experiment 2 investigated another kind of framing effect: loss aversion. The key statistic was an interaction between language and the size of a bet. Table S7 gives the *F* value. Experiment 3 explored willingness to take bets as a function of language. Table S7 gives the *t* (absolute) value.

Table S7: Statistical properties of the Keysar *et al.* (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1a	31, 30	6.14	.013	0.647	.715
Exp. 1a	30, 30	0.069	.793	--	.950
Exp. 1b	33, 33	10.88	.001	0.842	.928
Exp. 1b	39, 39	0.051	.821	--	.950
Exp. 1c	25, 25	5.30	.021	0.667	.655
Exp. 1c	26, 27	0.31	.578	--	.950
Exp. 1d	25, 25	26.1	<.001	--	.9996
Exp. 1d	28, 28	0.091	.763	--	.950
Exp. 2	68, 78	4.029	.047	0.331	.509
Exp. 3	27, 27	2.04	.046	0.547	.505
P_{TES}					.091

Keysar *et al.* (2012) reported that each of the studies produced a pattern of results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .091$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory. Note that the overall low power is largely due to Experiments 1a, 1c, 2, and 3. The strong findings in some of the other experiments suggest that the some of the empirical findings may be valid even though the set appears to be biased. However, Experiments 2 and 3 played an important role in the theory, so belief in some empirical results does not alleviate skepticism about the theory.

7. Analysis of Leung, Kim, Polman, Ong, Qiu, Goncalo, and Sanchez-Burks (2012) "Embodied Metaphors and Creative Acts"

Leung *et al.* (2012) reported five studies that purported to show a relationship between creative cognition and bodily experiences. Table S8 summarizes the statistics that contributed to the TES analysis.

Study 1 had participants perform a creativity task after switching an upheld hand or maintaining a held hand (to represent the metaphor "on the other hand"). The experiment measured several different variables and there were many different analyses. One key analysis was that participants in the switch condition were more creative than participants in the maintain condition. Table S8 reports the *t* value for this test. No tests on the other variables went against the theory, but it is not possible to estimate joint power for the full set of tests without knowing the correlation between variables. The power provided in Table S8 for this experiment is an upper limit of the joint power.

In Study 2a the key result was characterized by two hypothesis tests. An ANOVA reported a difference in creativity scores across conditions: inside-the-box, outside-the-box, and control. A planned contrast revealed that the outside-the-box scores were significantly higher than the other conditions. One can estimate the power of both tests to produce a significant result with simulated experiments. However, there appears to be a mismatch between the reported means and standard deviations and the reported *F* and *t* values. In particular, the reported standard deviations seem uncommonly small. Given the reported means and the *F* value, the pooled standard deviation is estimated to be 2.557, which is pretty close to what one would get if the reported standard deviations were actually standard errors (2.678). Using the 2.577 value, 100,000 simulated experiments drew samples from normal distributions with the given means and the pooled standard deviation, and power was estimated for finding significant results for both the main effect and the contrast.

Study 2b had a similar design to Study 2a, but two measures of creativity were reported for each participant. Both measures produced significant results, but the power of both measures cannot be calculated because the scores are correlated. The TES analysis focused on only the Lego task. Similar to Study 2a, the evidence for the theory was based on a main effect and a contrast. Simulations were again used to estimate the probability of both tests rejecting the null.

Study 3 found a significant difference between the creativity scores of two groups of participants who differed in how a virtual avatar could move. Table S8 reports the *F* value. Study 4 reported a significant difference in creativity scores for participants who first recombined split coasters compared to participants who simply moved coaster halves. Two recorded creativity measures were described as being related to divergent or convergent thinking. Statistics were only fully provided for the convergent measure (the divergent measure produced a predicted null result). Table S8 gives the reported *F* value.

Leung *et al.* (2012) reported that each of the studies produced a pattern of results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .076$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

Table S8: Statistical properties of the Leung *et al.* (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	20, 20	2.02	.050	0.626	.488
Study 2a	34, 34, 34	Multiple tests			.676
Study 2b	34, 35, 35	Multiple tests			.539
Study 3	36, 37	4.17	.045	0.473	.513
Study 4	32, 32	8.92	.004	0.738	.828
P_{TES}					.076

8. Analysis of Rounding, Lee, Jacobson and Ji (2012) "Religion and Self-Control"
Rounding *et al.* (2012) reported results from four experiments to support the theory that religious thoughts encourage self-control. Table S9 summarizes the statistics that contributed to the bias analysis. As evidence of improved self-control, Study 1 reported that participants in a religious-prime condition drank more unsavory juice than participants in a neutral-prime condition. Although it produced a statistically significant result, the experiment does not have high *post hoc* power.

Study 2 explored priming effects on delayed gratification. It reported that more participants in a religious prime condition were willing to wait in order to receive a larger payment. The test statistic in Table S9 is a χ^2 value. As in Study 1, the effect was statistically significant but the experiment was not very powerful. The effect size listed in Table S9 for Study 2 is Cohen's *h* (using n_1=28, p_1=0.607, n_2=32, p_2=0.344).

Study 3 examined the mechanisms of the religious-prime effect. The authors proposed a theoretical idea that priming refueled a limited resource of self-control. Study 3 measured the length of time subjects persisted in attempts to solve impossible puzzles. One group of participants experienced a neutral prime task. A second group of participants first participated in an ego-depletion task, which was predicted to reduce the persistence measure, before the neutral prime task. A third group of participants also participated in the ego-depletion task but then experienced a religious-prime task. Consistent with the theory, significant differences were found between the two neutral conditions and between the prime conditions that each underwent the ego-depletion task. Also consistent with the theory, the difference between the religious-prime condition and the neutral-prime without ego-depletion condition was not significant. All of these comparisons (both significant and non-significant) were important for validating the theory. Since there are various relationships between the tests, success probability was estimated as the proportion of times the multiple tests would show the full pattern of results that is consistent with the theory. The mean persistence times are provided in the text, and the pooled standard deviation (469.9) can be estimated from the standard errors in Figure 1 of Rounding *et al.* (2012). One hundred thousand simulated experiments were then generated. For each experiment the simulated data were tested for significance and success was indicated if the pattern of significant and non-significant results matched what Rounding *et al.* (2012) reported as being validation of the theory. This proportion is given in Table S9.

Study 4 also had multiple comparisons between four different priming conditions. Again the theory predicted a particular pattern: religious-priming should reduce reaction time on a Stroop task relative to a neutral task, but death-priming and morality-priming should not. The observed pattern of significant and non-significant findings was exactly as predicted. The sample sizes, means, and pooled standard deviation (0.074) were extracted from Figure 2 in Rounding *et al.* (2012). The probability of an experiment producing such a pattern of outcomes was estimated with 100,000 simulated experiments. The low probability indicated in Table S9 is because requiring a very specific pattern of results across multiple comparisons reduces the joint probability of that experimental outcome.

Table S9: Statistical properties of the Rounding *et al.* (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	23, 25	2.21	0.03	0.628	0.567
Study 2	28, 32	4.16	0.04	0.534	0.541
Study 3	19, 21, 20	Multiple tests			0.441
Study 4	24, 25, 20, 28	Multiple tests			0.268
P_{TES}					0.036

Every experiment in Rounding *et al.* (2012) produced the desired outcome. The probability of experiments like these producing such a pattern is the product of the probabilities: P_{TES} =.036. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

9. Analysis of Savani and Rattan (2012) "Choice and Inequality"

Savani and Rattan (2012) reported six experiments that purported to show that stimulating the concept of choice made people less disturbed by wealth inequality. Table S10 summarizes the statistics that contributed to the TES analysis.

Each experiment measured a wide range of participant variables and then used a regression analysis to explore whether participants in a condition that highlighted choice reported different attitudes about wealth than participants in a condition that did not highlight choice. Most of the statistics listed in Table S10 are *t* values. Experiment 5 was slightly different in that it used an ANCOVA to test a similar comparison, and the test statistic in Table S10 is an *F* value. Experiment 5 also included other measures, so the estimated power sets an upper limit.

It is a bit tricky to compute the standardized effect size for these kinds of designs because there are several choices for the standard deviation. In general, the analyses reported here used the standard deviation that was part of the reported test statistic (usually a *t* value). The effect size calculation also has to consider that the *t* statistic depends on not just the reported sample sizes but also the degrees of freedom, which is partly determined by the number of other variables in the linear regression or ANCOVA. Although these decisions can affect the reported effect size, they tend to make little difference (out to three

decimal places) in the final power calculations, which somewhat invert whatever process was used to generate the effect size.

In each experiment there were many different hypothesis tests that explored a variety of influences on judgments about wealth. Many of these tests were simply to control for factors (e.g., political orientation) other than the key comparison across conditions. Fortunately, the text in Savani and Rattan (2012) was very clear about which tests were important relative to the theoretical ideas.

Table S10: Statistical properties of the Savani and Rattan (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	24, 24	2.29	.027	0.650	.597
Exp. 2	23, 23	2.57	.014	0.745	.695
Exp. 3	34, 33	2.25	.028	0.543	.591
Exp. 4	30, 31	2.85	.006	0.721	.790
Exp. 5	35, 36	4.10	.047	0.475	.506
Exp. 6	25, 25	2.43	.019	0.677	.649
P_{TES}					.064

Savani and Rattan (2012) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .064$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

10. Analysis of van Boxtel and Koch (2012) "Visual Rivalry Without Spatial Conflict"

van Boxtel and Koch (2012) reported four experiments that explored visual rivalry in retinotopic and nonretinotopic reference frames. Table S11 summarizes the statistics that contributed to the TES analysis.

Each experiment measured visual rivalry for an apparent motion stimulus. In Experiment 1 participants made reports in each of four stimulus conditions. Support for the theoretical ideas was based on a statistical analysis that tested the reported bias for various stimulus conditions against fixed values (0.5 indicating maximum rivalry and 0 indicating no rivalry) and between select conditions. In total, Experiment 1 included eight different hypothesis tests; the theory predicted six significant results and two non-significant results. The text reports additional tests, but they seem to only be for completeness sake rather than being theory driven. The text also reports measures of perceptual clarity, but although these reports are related to the theory, there is no way to compute the joint power of both measures. The means and standard deviations could be estimated from the text and Figure 2 of van Boxtel and Koch (2012), although what are described as standard deviations in the text are actually standard errors. The probability of an experiment like this producing the full pattern of significant and non-significant results was estimated with 100,000 simulated

experiments having the mean and standard deviations indicated by the data. Even though many of the individual effects are quite strong, as Table S11 indicates, the probability of the full set is relatively low.

Experiment 2 examined whether variations to the stimuli in nonretinooptic and retinotopoic displays behaved similarly. The variations did produce significant changes to the rivalry reports, and they were similar for both conditions (no interaction). The within-subjects design of the experiment makes it difficult to estimate power for all of the hypothesis test outcomes. As an upper limit, the result listed in Table S11 is for just one of the tests, where the statistic is an *F* value.

Experiment 3 compared rivalry reports for stimulus conditions that produced group or element motion. The desired pattern of results involved four different hypothesis tests (two predicted significant and two predicted non-significant). To compute power it was necessary to estimate the correlation across participants for the reports in the two conditions. Applying algebra to the statistics for the reported hypothesis tests gives r= .240, and simulated data sets with such a correlation were created and then tested to estimate the probability of the entire pattern of results.

Experiment 4 had a design similar to Experiment 3 but with stimulus conditions that predicted no difference. The theory did still predict differences relative to fixed rivalry values. In all, there were two predicted significant results and three non-significant results. The correlation was estimated as r= .868. The success probability reported in Table S11 is low in part because many predicted non-significant results were sometimes significant for the simulated experiments.

Table S11: Statistical properties of the van Boxtel and Koch (2012) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	7	Multiple tests			.521
Exp. 2	7	8.7	.026	0.969	.575
Exp. 3	12	Multiple tests			.647
Exp. 4	6	Multiple tests			.364
P_{TES}					.071

van Boxtel and Koch (2012) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .071. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

11. Analysis of Evans, Horowitz and Wolfe (2011) "Weighting of Evidence in Rapid Scene Perception"

Evans *et al.* (2011) reported eight experiments that purported to show how observers rapidly extract information from visual scenes with multiple categories. Table S12

summarizes the statistics that contributed to the TES analysis. Each experiment measured accuracy or reaction time judgments about the presence or absence of one or more target categories. Table S12 reports the *t* values for these tests. In Experiment 1 a key comparison was that performance was above chance in a postcue condition.

The key finding in Experiment 2 was a predicted non-significant difference relative to chance. Table S12 reports the probability for the given effect size that a non-significant result would be produced. Experiment 3 was a characterization of effects across time, and the reported hypothesis tests were simply a description of effects and not directly related to the theoretical ideas. Experiment 4 reported a very strong effect across blocks that varied task relevance. Experiment 5 also investigated effects of task relevance with a reaction time measure.

Experiment 6 ruled out an alternative explanation for earlier results. Experiment 7 describes data but does not include a hypothesis test. Experiment 8 ruled out another alternative explanation.

Table S12: Statistical properties of the Evans *et al.* (2011) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	10	8.69	<.001	2.512	.999
Exp. 2	12	1.75	.110	0.470	.682
Exp. 3		No meaningful tests			
Exp. 4	20	12.83	<.001	2.754	1.00
Exp. 5	12	2.69	.021	0.722	.626
Exp. 6	20	5.43	<.001	1.166	.999
Exp. 7		No meaningful tests			
Exp. 8	12	12.26	<.001	3.29	1.00
P_{TES}					.426

Evans *et al.* (2011) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .426$. Since this value is larger than the .1 criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

12. Analysis of Inesi, Botti, Dubois, Rucker and Galinsky (2011) "Power of Choice"

Inesi *et al.* (2011) reported six experiments purporting to show that power (not in the statistical sense) and choice are substitutable. Table S13 summarizes the statistics, *t* values, that contributed to the TES analysis.

Experiment 1a reported that participants who recalled a low power personal experience preferred a larger set of consumer products (more choices) than participants who first recalled a high power personal experience.

Experiment 1b influenced power by having participants read a description of either a low-powered or high-powered individual. The experiment reported significant differences in measures of willingness to drive for a larger set of choices and for a willingness to wait for a larger set of choices. It might be possible to estimate the correlation between these measures and thereby compute the joint power, but the analysis is complicated by having some data points removed from the analysis (for being too extreme). Table S13 reports the statistics only for the driving measure, which is an upper limit of the joint power for both measures.

Experiment 2a examined how choice influenced desire for power. The design of the experiment is complicated, and not enough information is provided to compute power for many of the various hypothesis tests. Table S13 reports a *t* value for one of the tests (low-choice participants gave more importance to power-relevant features than high-choice participants). The statistical power listed in Table S13 is an upper limit for the joint power of the full set of results.

Experiment 2b investigated a similar relationship by manipulating product status and asking participants how much they would be willing to pay. The data produced a significant interaction between product status and choice conditions. Payment for low status products was unaffected by the choice condition, but participants in the small choice condition were willing to pay more for high status products. Both of these results were important for the theoretical claims. There was no choice difference for low-status products, but it is not clear that the absence of this effect is predicted by the theory. One hundred thousand simulated experiments used the means and standard deviations in Inesi *et al.* (2011) to estimate the probability of rejecting both the interaction and the effect of choice for high-status products. Experiment 2b also included a measure of feeling personal control, which was shown to mediate the other measures. It does not appear to be possible to include these effects in the power calculation. The power given in Table S13 is an upper limit of the power for all the reported effects.

Experiments 3a and 3b tested whether choice and power contributed similarly to feelings of personal control. Each experiment used a 2 × 2 between subjects design that varied power and choice. Relevant to the theoretical claims, Experiment 3a reported a main effect of choice, a significant interaction, a significant contrast for participants without power or choice compared to participants with either one, and a non-significant contrast for participants with either power or choice compared to participants with both. In a break from convention, statistical significance for the interaction was claimed with p=.054, which suggests a criterion of .055. Using the means and standard deviations provided in the text (Table S5), simulated experiments were generated to estimate the probability of this kind of experiment satisfying all of the theoretically motivated outcomes. These simulations used the non-standard .055 criterion for judging significance of the interaction. Table S13 indicates that the joint probability is quite low.

Experiment 3b had a similar design and analysis. Power was estimated in a similar way. Consistent with the original analysis, the traditional .05 criterion was used to judge significance for the interaction.

Table S13: Statistical properties of the Inesi *et al.* (2011) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1a	20, 20	2.14	.039	0.663	.534
Exp. 1b	20, 20	2.08	.044	0.645	.511
Exp. 2a	26, 26	2.34	.023	0.639	.618
Exp. 2b	15, 15, 15, 15	Multiple tests			.625
Exp. 3a	27, 28, 27, 28	Multiple tests			.350
Exp. 3b	22, 22, 21, 21	Multiple tests			.716
P_{TES}					.026

Inesi *et al.* (2011) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .026$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

13. Analysis of Nordgren, Morris McDonnell and Lowenstein (2011) "What Constitutes Torture?"

Nordgren *et al.* (2011) reported four experiments purporting to show that people were more likely to classify an interrogation tactic as torture if they experienced a mild version of the specific induced pain. Table S14 summarizes the power values that contributed to the TES analysis. The table includes few specific statistics because every study included multiple hypothesis tests.

Study 1 included two behavioral measures for people in three conditions: social pain, no-pain, and control. Participants were asked to estimate the pain of being in solitary confinement and to indicate whether they oppose using solitary confinement in jails. The correlation between the two measures was not provided, so power can only be estimated for one of the measures. The second measure had the smaller significant effect, so its power establishes a tighter upper bound on estimated power. The statistical test was a χ^2 test for three proportions across the conditions. The power of the test was estimated with 100,000 simulated experiments that produced samples from populations having the proportions indicated in the study (p_1=.630, p_2=.333, p_3=.357). The χ^2 value given in Nordgren *et al.* (2011) appears to be slightly off, and the correct value is given in Table S14. This difference does not change the (rounded) reported *p* value of the test.

Study 2 examined the specificity of the experience. Participants were either fatigued or not and judged the ethics of either sleep deprivation or cold treatment. There was a predicted significant interaction, a predicted effect of fatigued participants judging sleep deprivation to be less ethical than nonfatigued participants, and a non-significant effect for judgments of cold treatment. The participants also evaluated the pain of these treatments, but this measure is correlated with the ethics evaluation, so it is only possible to estimate the success probability for one of the variables. From the interaction *F* value and the

reported means, it is possible to estimate the pooled standard deviation (1.088) and then generate 100,000 samples using the reported means. Each sample was tested for the desired pattern of test outcomes, and the joint success probability of the set is the proportion of samples that match all three outcomes.

Study 3 assigned participants to one of three conditions (cold, prior cold, control) where the coldness conditions required participants to hold their arm in a bucket of cold water. Participants then rated the pain of an interrogation process that used coldness and also evaluated whether such a process was torture. A significant difference was found between participants currently holding their arm in cold water (the cold condition) and the control condition and also between the cold and prior cold (participants had their arm in cold water 10 minutes prior to the judgments) conditions. Using the means and standard deviations reported in the text for the torture judgments, 100,000 simulated experiments generated random samples and repeated the hypothesis tests. The power analysis does not include the pain ratings, so it only provides an upper limit on the power of both results.

Study 4 dropped the prior cold condition and included a cold weather condition. Again, two measures were reported and both showed a main effect of condition, a significant contrast for participants in the cold weather and control conditions, and no significant difference between participants in the cold condition and the cold weather condition. It is not possible to compute the power for tests of both measures without knowing the correlation, so the success probability in Table S14 is for only one of the variables (support for cold weather as punishment). It was estimated by using the reported means and pooled standard deviation (0.987), which was estimated from the provided *F* value. It is worth noting that the reported standard deviations for these scores do not agree with the reported means and *F* values. The analysis included an ANOVA and two follow-up tests. The follow-up tests could have been performed as contrasts, which would take advantage of a more precise standard deviation and larger degrees of freedom, but instead Nordgren *et al.* appeared to run independent tests. The simulations were run the same way.

Table S14: Statistical properties of the Nordgren *et al.* (2011) experiments.

	n	**Test Statistic**	**p**	**Effect size**	**Probability of Success**
Study 1	27, 33, 28	6.25	.044	--	.613
Study 2	27, 27, 27, 27	Multiple tests			.396
Study 3	24, 24, 25	Multiple tests			.456
Study 4	22, 22, 23	Multiple tests			.812
P_{TES}					.090

Nordgren *et al.* (2011) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .090$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

14. Analysis of Savani, Stephens and Markus (2011) "Interpersonal and Societal Consequences of Choice"

Savani *et al.* (2011) reported five experiments purporting to show that people activating the concept of choice have decreased support for equality and related ideas. Table S15 summarizes the power values that contributed to the TES analysis.

In Study 1, participants answered a short survey about affirmative action. In a choice condition participants first watched a movie and indicated whenever a character made a choice. Control participants watched the movie without the choice classification task. A *t* test revealed a significant difference between support for affirmative action, with those in the choice condition giving less support. The reported *p* value (.05) is actually rounded down from .055. Even supposing that the *t* value is rounded down does not reach the standard criterion *p* value of .05. To be consistent with the apparent criterion in Savani *et al.* (2011), the power calculation used a criterion of .055.

Study 2 had a similar design and used similar methods while comparing a wide range of societal problems. A significant result was found with a *t* test using the .05 criterion.

Study 3 contrasted support for policies that expanded or restricted individual rights. Participants in a choice condition reported greater support for policies that expanded individual rights, while participants in the control condition supported the types of policies equally. Since the participants were part of separate samples, the conditions were tested with separate within-subject *t* tests. Table S15 reports the properties of both tests. The effect sizes in Table S15 use the standard deviation of the difference scores, which allows for direct calculation of power but differs from what is reported in Savani *et al.* (2011). For Study 3b (controls), Table S15 reports the probability of not rejecting the null hypothesis, as predicted by the theory.

Study 4 reported a *t* test for a significant factor of condition (choice or control) in a regression. The estimated effect size does not properly take into account the reduction in degrees of freedom due to other regression factors, but the power value ends up being little changed by this approximation. Study 5 was a cross-culture study that compared choice and control conditions for Americans and Indians. Americans in the choice condition had significantly less empathy for a poor child, but Indians (Study 5b) did not differ by condition. The success probability for Study 5b in Table S15 is for not finding a significant difference for the Indians.

Savani *et al.* (2011) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .063$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

Table S15: Statistical properties of the Savani *et al.* (2011) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	25, 25	1.97	.055	0.548	.493
Study 2	17, 18	2.21	.034	0.730	.554
Study 3a	47	3.82	<.001	0.548	.957
Study 3b	51	0.52	.605	0.072	.921
Study 4	27, 27	2.12	.039	0.569	.536
Study 5a	13, 13	2.25	.034	0.855	.552
Study 5b	23, 24	0.78	.439	0.224	.883
P_{TES}					.063

15. Analysis of Todd, Hanko, Galinsky and Mussweiler (2011) "Difference Mind-Set and Perspective Taking"

Todd *et al.* (2011) reported five experiments purporting to show that a difference mind set facilities perspective taking. Table S16 summarizes the statistical values that contributed to the TES analysis.

In Experiment 1 participants were assigned to three conditions that introduced a similarity-mind-set, a difference-mind-set, or a control situation. Consistent with the theory, the difference condition was significantly different from both of the other conditions, which were not significantly different from each other. The probability of the entire set of outcomes was estimated with 100,000 simulated experiments that were defined by the sample means and standard deviations. Experiment 2 had similar conditions and hypotheses as Experiment 1 (but a different behavioral measure), and success probability was estimated in a similar way.

Experiment 3 included only the similarity and difference conditions to study their effects on use of privileged knowledge. Consistent with the prediction, participants in the difference condition were less likely to use their privileged knowledge to infer the actions of a character in a story. The statistic in Table S16 is a *t* value.

Experiment 4 was similar to Experiment 3, but varied whether the story character had an ethnicity that promoted in-group or out-group categorization. As predicted, there was a significant difference in use of privileged knowledge across categorizations. The statistic in Table S16 is a *t* value.

Experiment 5 measured maze completion time as a function of in-group and out-group categorizations in an interpersonal coordination task. As predicted, times were shorter for out-group pairs. The statistic in Table S16 is an *F* value, and the sample size indicates the number of pairs (each pair contributed one score).

Todd *et al.* (2011) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .043$. Since this value is smaller than the .1

criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

Table S16: Statistical properties of the Todd *et al.* (2011) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	27, 27, 27	Multiple tests			.482
Exp. 2	33, 33, 34	Multiple tests			.522
Exp. 3	20, 20	2.27	.029	0.704	.583
Exp. 4	29, 29	2.20	.032	0.570	.569
Exp. 5	14, 14	4.61	.041	0.788	.519
P_{TES}					.043

16. Analysis of Tuk, Trampe, and Warlop (2011) "Inhibitory Spillover"

Tuk *et al.* (2011) reported four experiments purporting to show a connection between bladder control and the ability to exert self-control. Table S17 summarizes the statistical values that contributed to the TES analysis.

In Study 1 participants completed a Stroop task and reported their need to urinate urgency. A regression showed that response times decreased for larger urgency reports. The statistic in Table S17 gives the *t* value. This experiment also measured responses times for a non-interfering version of the Stroop task and, as expected, found no relationship with urination urgency. Since there is a correlation between these measures, it is not possible to measure the joint probability of these outcomes. Thus, the power in Table S17 provides an upper limit for the power of the joint effects.

Study 2 manipulated bladder pressure by having participants sip from or drink all of five cups of water. Participants then completed a delayed reward task. Participants in the high-bladder-pressure condition showed increased impulse control. The statistic in Table S17 is a *t* value. Study 3 replicated Study 2; it also included some moderation analyses, but it is not possible to estimate the joint power.

Study 4 primed half the participants to think about urination by giving them a word search task with urination-related words; the other half of the participants, a control condition, searched for other words. The behavioral measure was the same delayed reward task used previously. Across priming conditions, participants reported statistically significant differences in urination urgency. More importantly, these reported urgencies mediated performance on the delayed reward task. The latter result is the key finding, and its power was estimated by examining the 95% confidence interval, which is given in Table S17 and indicates statistical significance by not containing zero. If the distribution of the mediation effect follows a normal distribution with mean 0.236 and standard deviation 0.120, then repeating its construction for different samples would mean it would not contain zero 52% of the time (power). Even if the normal distribution is not quite right, the power is unlikely to be much different from the estimated value.

Table S17: Statistical properties of the Tuk *et al.* (2011) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	88, 88	2.13	.035	0.320	.559
Study 2	51, 51	2.20	.030	0.432	.580
Study 3	48, 49	2.11	.037	0.425	.545
Study 4	65, 66	[0.006, 0.49]	.050	--	.520
P_{TES}					.092

Tuk *et al.* (2011) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .092$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

17. Analysis of Balcetis and Dunning (2010) "Wishful Seeing"

Balcetis and Dunning (2010) reported five experiments purporting to show that object distance estimates were influenced by the desirability of the object. Table S18 summarizes the statistical values that contributed to the TES analysis. All statistics are *t* values. Francis (2012b) previously described this TES analysis. In a reply, Balcetis and Dunning (2012) explained that a non-significant finding was removed from the original article at the request of reviewers.

Every study measured a distance estimate for two different conditions. Although the condition inducers and measurement methods differed somewhat across conditions, the effect sizes are quite similar relative to their uncertainty (response in Francis, 2012b), so pooling the effect sizes with meta-analytic techniques provides the best estimate of experimental power. The pooled effect size is estimated to be 0.537.

Study 1 had thirsty or quenched participants estimate the distance of a bottle of water. The estimates were smaller for thirsty participants. In Study 2a participants estimated the distance of a $100 bill. This estimate was smaller when the participants had a chance to win the bill. In Study 2b participants completed a sense of humor survey and then estimated its distance. The estimates depended on whether the survey had been scored positively or negatively.

Study 3a had participants throw a beanbag to a gift card that either had value or not. Throws were shorter when the card had desirable value. Study 3b had participants estimate the distance (using a distance matching procedure) to either chocolate or dog feces. The estimated distances were smaller for chocolate.

All of these experiments also included several other measures about mood and attitude. Apparently these measures were to rule out other possible effects that would influence the estimates. It is not clear that they are important for the theoretical claims, but if they were important then the power values estimated in Table S18 would be even smaller.

Table S18: Statistical properties of the Balcetis and Dunning (2010) experiments. Power is estimated with the pooled effect size of 0.537.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	47, 43	2.00	.049	0.418	.712
Study 2a	60, 61	2.84	.005	0.513	.834
Study 2b	42, 47	2.01	.048	0.423	.706
Study 3a	20, 20	3.43	.001	1.063	.381
Study 3b	26, 26	2.29	.026	0.626	.476
P_{TES}					.076

Balcetis and Dunning (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .076$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

18. Analysis of Bowles and Gelfand (2010) "Status and Workplace Deviance"

Bowles and Gelfand (2010) reported four experiments purporting to show that the relative status of an evaluator influences both rule observance and attitudes toward social hierarchy. Table S19 summarizes the statistical values that contributed to the TES analysis.

In Study 1 participants were assigned to high or low status conditions by imagining situations where others looked up to them or they looked up to others. A survey then measured rule compliance and social dominance. There were differences between conditions for both of these measures, but since they are correlated the joint power for both tests cannot be estimated. An upper estimate is found by estimating the power of one of the measures (social dominance). The reported test statistic in Table S19 is a *t* value.

Study 2a manipulated participant status using the same technique and then evaluated whether to punish a low or high status target employee. A between subjects ANOVA reported no effect of evaluator status, a significant effect of target status, and a significant interaction. Subsequent *t* tests also found that high-status evaluators were more inclined to punish a low-status target than a high-status target but that there was no difference in targets for the low-status evaluators. All of these results were deemed important for the theoretical claims. Study 2b was a replication of 2a, with a slight modification of the target status, and used the same kind of analysis. For both of these studies, success probability was estimated from 100,000 simulated experiments.

Study 3 investigated whether (White) men and women evaluated others in a similar way. The study investigated several different behaviors and conditions, but it is not possible to compute the power for all of the possibilities without knowing the correlation between effects. As an upper estimate, Table S19 reports the *F* value for White men who evaluated male deviance more leniently than female deviance (labeled as Study 3a). In contrast,

White females (labeled as Study 3b) showed no significant difference. The text in Bowles and Gelfand (2010) does not give precise summary statistics for this measure, but the probability of not rejecting a true null effect would be .95.

Table S19: Statistical properties of the Bowles and Gelfand (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	107, 108	2.75	.006	0.374	.779
Study 2a	73, 73, 67, 68	Multiple tests			.489
Study 2b	43, 44, 44, 43	Multiple tests			.241
Study 3a	21, 22	6.04	.018	0.736	.654
Study 3b	58, 58	<0.04			.950
P_{TES}					.057

Bowles and Gelfand (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .057$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

19. Analysis of Damisch, Stoberock and Mussweiler (2010) "How Superstition Improves Performance"

Damisch *et al.* (2010) reported four experiments purporting to show that activating good-luck superstitions improved performance in a variety of tasks. Table S20 summarizes the statistical values that contributed to the TES analysis.

In Study 1 participants given what was described as a "lucky ball" demonstrated better putting performance than a control group. The reported test statistic in Table S20 is a *t* value.

Study 2 included one experimental condition where participants were wished good luck before completing a motor-dexterity task and two control conditions. The analysis involved an ANOVA and subsequent contrasts that showed significant differences between the experimental condition and each control condition, but no difference between the control conditions. It should be mentioned that the reported *F* value for the ANOVA appears to be too small. It does not agree with the reported means and standard deviations or with the reported *p* value for the test. The correct *F* value should be 3.48. The success probability in Table S20 is estimated from 100,000 simulated experiments that drew samples from populations having the reported means and standard deviations.

Study 3 had participants perform a memory task either with or without a personal lucky charm. Performance was better for participants with their lucky charm. The statistic in Table S20 is a *t* value. Study 4 used similar methods as Study 3 but measured performance on an anagram task. Both of these studies included several other measures as well, but the joint power cannot be estimated without knowing the correlation between

variables; thus the power values provided in Table S20 establish upper limits for the joint power.

Table S20: Statistical properties of the Damisch *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	14, 14	2.14	.042	0.785	.516
Study 2	17, 17, 17	Multiple tests			.415
Study 3	21, 20	2.13	.040	0.653	.531
Study 4	15, 14	2.10	.045	0.759	.504
P_{TES}					.057

Damisch *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .057$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

20. Analysis of de Hevia and Spelke (2010) "Number-Space Mapping in Human Infants"

de Hevia and Spelke (2010) reported four experiments purporting to show that preverbal infants were sensitive to an ordered series of numerosities. Table S21 summarizes the statistical values that contributed to the TES analysis.

In Experiment 1 infant participants habituated to visual dot displays that systematically changed in numerosity. During subsequent test trials, the infants were presented with sequences of lines that systematically changed in numerosity. The dependent variable was looking time. A *t* test found that infants spent more time looking at the lines that changed differently than the habituation trials. As an additional analysis, a sign test found that 18 of 24 infants showed the novelty preference. Both tests were part of the power estimation. In simulated experiments the difference scores for each participant were sampled from a normal distribution having a mean that reflected the standardized difference of scores to be consistent with the *t* test. These scores were then analyzed for the *t* test and sign test. Success for the experiment required both tests to show a significant result. The sign test has rather low power (0.31) by itself, because it would be non-significant if only 17 of 24 infants showed a novelty preference. The joint power of both tests is only slightly lower. Experiment 2 was a replication of Experiment 1, and power was estimated in the same way. Here the sign test is more powerful by itself (0.61) and the joint power is only slightly smaller.

Experiment 3 set up stimulus sequences that were predicted to produce non-significant effects. There were several predicted null findings, but to compute their joint probability would require knowing the correlation between measures. As an upper limit,

Table S21 computes the probability for the *t* test that has the lowest probability of producing a non-significant result.

Experiment 4 was a modification of Experiments 2 and 3 that used within-subjects ANOVAs across three test-trials. The nature of the analyses prohibits a direct calculation of power for any of the ANOVAs because the correlation between measures was not given. An approximate upper bound for the joint power can be given by noting that one of the tests produced a *p* value that was exactly at the significance criterion (.05). Typically, replications are equally likely to produce *p* values above or below the criterion, so the power is 0.5. This power value is an upper limit of the joint power across the many hypothesis tests in this experiment.

Table S21: Statistical properties of the de Hevia and Spelke (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	24	Multiple tests			.297
Exp. 2	20	Multiple tests			.591
Exp. 3	20	-1.24	.231	-0.266	.795
Exp. 4	20	--	.05	--	.500
P_{TES}					.070

de Hevia and Spelke (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .070$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

21. Analysis of Ersner-Hershfield, Galinsky, Kray and King (2010) "Counterfactual Reflection"

Ersner-Hershfield *et al.* (2010) reported four experiments purporting to show that thinking about counterfactual origins (how things might have turned out differently) increased commitment to various organizations compared to factual origin thinking (describing how things happened). Table S22 summarizes the statistical values that contributed to the TES analysis.

Study 1 measured patriotic attitudes for participants who first engaged in either factual or counterfactual thinking about their home country's origins. As predicted, participants in the counterfactual condition reported significantly higher patriotism. The statistic in Table S22 is a *t* value. The reported *p* value (.053) is above the typical criterion for significance, so the power calculation used a criterion of .053.

Study 2 was similar to Study 1 but measured commitment to a company. The statistic in Table S22 is a *t* value, and again the *p* value (.056) is larger than the typical criterion. The power analysis used a criterion of .057. Study 2 also measured other

variables to perform a mediation analysis and compared judgments of company trajectory. The power in Table S22 provides an upper limit on the joint power of the analyses.

Study 3 manipulated type of thinking for the origins of a fictional company. A significant difference in commitment was reported, and the statistic in Table S22 is an *F* value. This study varied the type of company, but this was not central to the theoretical ideas. The study also explored mediation with several variables, so the power value reported in Table S22 is an upper limit on the joint power of the analyses.

The design of Study 4 is rather complex. The hypothesis test ended up comparing the probability a participant would be more likely to email a contact that they had previously thought about factually or counterfactually. The hypothesis test used Cochran's *Q*, which is compared to a χ^2 critical value. The power of the test was estimated with an on-line power calculator (Lenth, 2009).

Table S22: Statistical properties of the Ersner-Hershfield *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	20, 20	2.00	.053	0.620	.491
Study 2	48, 49	1.93	.056	0.389	.497
Study 3	30, 30	4.15	.046	0.519	.530
Study 4	21	4.50	.034	--	.564
P_{TES}					.073

Ersner-Hershfield *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .073$. Since this value is smaller than the .1 criterion, readers should be skeptical about the reported experimental results as they relate to the theory.

22. Analysis of Gao, McCarthy and Scholl (2010) "The Wolfpack Effect"

Gao *et al.* (2010) reported seven experiments describing properties of randomly moving shapes that all point toward a target. Table S23 summarizes the statistical values that contributed to the TES analysis.

Experiment 1 compared the wolfpack effect to other cues for animacy by measuring how different stimulus conditions interfered with detection of pursuit. The wolfpack effect introduced significantly more interference than two other conditions. The measures were from a single sample of participants, so measuring power of the joint effects required estimating the correlations between the measures. The correlations were calculated by computing the standard deviation of the differences (from the reported *t* values) and combining them with the standard deviation of each condition alone. Algebra can solve for the correlations, which are provided in the on-line Excel files. The success probability given in Table S23 is the probability of the wolfpack effect being significantly different from the perpendicular, match, and disc conditions. It was estimated from 100,000

simulated experiments with conditions having the means, standard deviations, and correlations reported in the original study.

In Experiment 2 participants moved an on-screen element to try to avoid a single pursuer. The task was more difficult when the pursuer was embedded in a wolfpack display than in a perpendicular display. The statistic in Table S23 is a *t* value.

Experiments 3a-d explored the aversive nature of the wolfpack display in various forms. Participants moved a screen element around to avoid contact with other elements. The screen was constructed to have some regions with a wolfpack type display and some regions without it. The measure was the amount of time spent in different regions. Experiments 3c and d had stimulus conditions that predicted a null result. The success probabilty in Table S23 is the probability of getting a null result.

Experiment 4 was like Experiment 3, but there were three display types. Sometimes the wolfpack focused on the user-controlled element, other times the wolfpack elements were oriented orthogonal to the user's element, and some times the wolfpack elements focused on the wolf. Performance in the wolfpack-to-wolf condition was better than the other conditions. Table S23 gives the *F* value for comparing the wolfpack-to-wolf and perpendicular-to-sheep conditions. Other hypothesis tests were much stronger, so this establishes the tightest upper bound on the joint power. The effect size and power estimation is complicated because the analysis is a contrast that pools the standard deviation across all conditions and thereby changes the degrees of freedom. The effect size in Table S23 uses n=10 to compute the effect size, based on the standard deviation of difference scores, and to then compute power. This effect size is not useful outside of the power calculation, and the power calculation itself may be less sure than for other conditions.

Gao *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .115. Since this value is larger than the .1 criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

Table S23: Statistical properties of the Gao *et al.* (2010) experiments.

	n	**Test Statistic**	**p**	**Effect size**	**Probability of Success**
Exp. 1	12	Multiple tests			.418
Exp. 2	8	3.03	.019	0.952	.639
Exp. 3a	10	4.18	.002	1.210	.924
Exp. 3b	7	5.29	.002	1.739	.967
Exp. 3c	14	0.67	.515	0.169	.910
Exp. 3d	9	0.71	.499	0.213	.913
Exp. 4	12	7.08	.026	0.769	.583
P_{TES}					.115

23. Analysis of Lammers, Stapel and Galinsky (2010) "Power and Hypocrisy"

Lammers *et al.* (2010) reported five experiments purporting to show that power increases moral hypocrisy. Table S24 summarizes the statistical values that contributed to the TES analysis. It should be noted that the Levelt, Noort & Drenth Committees (2012), which investigated data fraud by Diederik Stapel, did not identify this work as containing data fraud.

Experiment 1 reported that participants in a high-power condition (invoked by recalling a high-power experience) were both less accepting of overreporting travel expenses and more likely to cheat on reporting the outcome of dice rolls (which gained lottery tickets) compared to participants in a low-power condition. The analysis identified a significant interaction between power and task, a significant difference between high and low-power participants for the acceptance judgment, and a significant difference between high and low-power participants for the dice roll reports. The analysis was complicated in two ways, so success probability was computed with simulated experiments. The first complication was that, to enable comparisons across the judgment (acceptance of overreporting) and behavior (report on dice rolls) measurements, the scores were standardized to avoid differences in scale. The text is not clear on how this was done, but a reasonable approach is to rescale both measurements to a scale between 0 and 1. A few other possibilities were also considered, but they made little difference in terms of success probability calculations. The second complication was that statistical significance was claimed for the judgment task with p=.08, which is above the traditional .05 criterion. To estimate success probability, the simulations concluded significance for a p value below .1. Success probability was estimated from 100,000 simulated experiments that drew samples from populations matching the means and standard deviations of the samples. Each sample was then subject to the same analysis as the original data, and success required all three tests to show a significant result. The resulting success probability is quite low, possibly because, for many data sets, the standardization and subsequent pooling of the standard deviation make it less likely that the test will reject the null hypothesis.

Experiments 2, 3 and 4 avoided the scale issue problems of Experiment 1 by having all participants read a moral dilemma. Participants in one group then rated the acceptability of other people to break norms relative to the dilemma. Participants in another group rated the acceptability of themselves to break norms. There was again a manipulation of power, so each experiment was a 2 × 2 design. In each experiment, the key result was that, for high-power participants, it was deemed more appropriate for themselves to break norms than for others to break norms (hypocrisy). Table S24 reports the t values, which were based on contrasts across all groups, for these comparisons. The analysis also tested for differences among the low-power participants, but the outcome was deemed unrelated to the theory. These latter findings do not contribute to the success probability calculations. Since the tests were contrasts that pooled standard deviations, the success probability for each experiment was estimated with 100,000 simulated experiments. For Experiment 4, significance was reported for p=.06, so that was used as the criterion in the success probability calculations for that experiment. There were additional hypothesis tests that pooled the findings across Experiments 2, 3, and 4. Adding those effects to the success

probability calculations would likely only slightly reduce the joint success probability across experiments.

Experiment 5 used a similar design, but included conditions with legitimate and illegitimate power. The theory predicted, and found, a significant difference for legitimate power but not for illegitimate power. Significance was concluded with a contrast and used a non-standard *p* value (.0645). For illegitimate power a significant effect was found in the opposite direction but this was neither predicted nor required by the theory. The success probability calculation in Table S24 is based only on the findings for the legitimate high-power condition. The analysis also included some other comparisons, but it is not clear whether they were necessary for the theoretical conclusions. Success probability was estimated from 100,000 simulated experiments using a significance criterion of .065.

Table S24: Statistical properties of the Lammers *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	15, 15, 15, 16	Multiple tests			.179
Exp. 2	10, 11, 10, 11	2.40	.021	--	.703
Exp. 3	22, 22, 22, 22	2.21	.030	--	.628
Exp. 4	10, 11, 10, 11	1.94	.060	--	.508
Exp. 5	13, 13, 13, 13 13, 13, 13, 14	1.87	.065	--	.595
P_{TES}					.024

Lammers *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .024. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

24. Analysis of Li, Wei and Soman (2010) "Physical Enclosure and Psychological Closure"

Li *et al.* (2010) reported four experiments purporting to show that negative emotions were alleviated by the physical act of enclosing an emotional stimulus. Table S25 summarizes the statistical values that contributed to the TES analysis.

In Experiment 1a participants wrote about a past decision that they regretted and then provided self-emotion ratings. Participants who placed their writing in an envelope reported feeling less negative than participants who did not use an envelope. Experiment 1b used a similar design, but asked participants to write about an unfulfilled desire and measured mental disturbance. Table S25 lists the *t* values used for comparing these groups. Experiment 1b also asked participants about negative emotions, but the two measures are correlated and it is not possible to compute the joint power. The power value in Table S25 is an upper limit.

In Experiment 2 participants did not use an envelope, placed emotional content in an envelope, or placed unrelated items in an envelope. The analysis involved four contrasts that involved two significant findings, one marginally significant finding, and one predicted null result. The success probability was estimated with 100,000 simulated experiments that drew samples from populations having the means and standard deviations described by the reported data. The success probability reported in Table S3 is the proportion of times a set of samples produced the desired pattern of results.

Experiment 3 had participants either enclose their writing or clip the pages together. Both negative emotions and psychological closure were different across the groups. The correlation between the scores can be estimated by a regression of felt closure on emotion intensity (r=.505) and thereby allow the joint success probability for both effects to be estimated with 100,000 simulated experiments.

Table S25: Statistical properties of the Li *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1a	40, 40	2.24	.028	0.496	.591
Exp. 1b	20, 20	2.72	.009	0.843	.738
Exp. 2	27, 19, 34	Multiple tests			.382
Exp. 3	23, 24	Multiple tests			.476
P_{TES}					.079

Li *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .079. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

25. Analysis of Maddux, Yang, Falk, Adam, Adair, Endo, Carmon and Heine (2010) "Culture and the Endowment Effect"

Maddux *et al.* (2010) reported four experiments purporting to show that East Asians produced a smaller endowment effect (owners value objects more than potential buyers) than Westerners. Table S26 summarizes the statistical values that contributed to the TES analysis. For all of the experiments the analysis included multiple significant effects from an ANOVA. The success probability values are estimated with 100,000 simulated experiments that drew samples from populations having the means and standard deviations of the original samples.

Study 1a measured the endowment effect for European Americans and Asian Americans. The analysis identified significant overall endowment effects for European Americans and Asian Americans, separately, and an interaction that showed a significant difference in the endowment effect across the groups. Study 1b used a similar design with Canadians.

Study 2 divided participants into three groups: control, independent, and interdependent, where the latter two conditions involved priming for independent or interdependent thinking by writing an essay. The analysis used an ANOVA to find a significant endowment effect and an interaction with condition. An additional five contrasts compared the endowment effect within and across conditions.

Study 3 had Japanese and Canadian participants write an essay about an object that was either important or unimportant to them, and then evaluate the value of the object as either a buyer or a seller. The analysis included six hypothesis tests. These included a main effect of role (buyer or seller), a three-way interaction between culture, role, and writing task, and then contrast comparisons within culture or between cultures. The joint success probability of these outcomes included one predicted null result and one marginal result.

Table S26: Statistical properties of the Maddux *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1a	29, 30, 28,27	Multiple tests			.446
Study 1b	22, 23, 30, 30	Multiple tests			.599
Study 2	16, 16, 15, 16, 16, 16	Multiple tests			.234
Study 3	19, 19, 19, 20 14, 14, 14, 14	Multiple tests			.221
P_{TES}					.014

Maddux *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .014$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

26. Analysis of McGraw and Warren (2010) "Benign Violations"

McGraw and Warren (2010) reported five experiments purporting to show that amusement is related to violations of norms that are seen as benign. Table S27 summarizes the statistical values that contributed to the TES analysis.

Study 1 had participants evaluate stories that either benignly or morally violated a norm; moral violations were more often judged as amusing/wrong compared to stories that do not violate a moral norm. The analysis included three hypothesis tests, and they all showed such large effects that the power is effectively one. Table S27 gives the smallest *F* value.

Study 2 asked participants to judge whether behavior in a story could be judged as both wrong and not wrong, depending on the situation. The researchers also noted signs of amusement. The results were analyzed by comparing the proportion of participants who showed signs of amusement while reading a story with a violation rather than a control story. The statistic in Table S27 is a χ^2 value. Another analysis compared the proportions of

participants showing amusement according to whether it was also judged strictly wrong or strictly not wrong. These two results are related to each other in unclear ways. Power was estimated by the weaker effect, which sets an upper limit on the power of the joint outcomes.

Studies 3, 4, and 5 explored whether benign moral violations elicited mixed emotions of amusement and disgust. Study 3 compared reports of amusement for participants who read either a harmless or harmful version of a story. The harmless version produced more amusement. The same participants also reported disgust, with a significant difference in both amusement and disgust reports across story versions. As it turns out, these two tests were almost identical because almost every participant reported the story (both versions) to be disgusting. The joint power in Table S27 was estimated with 100,000 simulated experiments.

Study 4 had a similar design as Study 3, but instead of various stories, the key comparison was for participants who attended church versus those who did not. There were also variations in the stories, but these were not central to the theory (a predicted, and found, null result). Again, separate tests were run to examine differences in the frequency of reporting just amusement and of reporting both amusement and disgust. The success probability in Table S27 is the joint power for both tests, but does not include the predicted null finding.

Study 5 also had a similar design but varied "psychological distance" by having participants draw a plot with near or far points. Differences between these groups were found for judgments of amusement and both amusement and disgust. The success probability in Table S27 is the joint power for both tests, and it was estimated with 100,000 simulated experiments.

Table S27: Statistical properties of the McGraw and Warren (2010) experiments.

	n	**Test Statistic**	**p**	**Effect size**	**Probability of Success**
Study 1	33, 33	60.7	<.001	1.90	1.00
Study 2	36, 37	6.5	.010	--	.773
Study 3	18, 18	Multiple tests			.507
Study 4	26, 13	Multiple tests			.402
Study 5	22, 18	Multiple tests			.511
P_{TES}					.081

McGraw and Warren (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .081$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

27. Analysis of Sackett, Meyvis, Nelson, Converse and Sackett (2010) "When Time Flies"

Sackett *et al.* (2010) reported eight experiments purporting to show that perceived time distortion alters enjoyment of an experience. Table S28 summarizes the statistical values that contributed to the TES analysis.

Study 1a reported a *t* test that compared enjoyment ratings for participants who spent more (time-drags) or less (time-flies) time than expected on a task. Study 1b had a similar design and analysis.

Study 2 found that participants in a time-flies condition reported less negativity for an annoying song than participants in a time-drags condition (both conditions listened for the same length of physical time). The statistic in Table S28 is an *F* value. There was an additional analysis for another dependent variable, but the joint power cannot be computed with the provided information. The power value in Table S28 is an estimated upper limit.

In Study 3 participants selected a favorite song and listened to it while a timer counted up or down and also varied in whether it accelerated (time-flies) or decelerated (time-drags). The analysis reported a significant interaction between timer speed and direction and a contrast showing significantly more enjoyment in the time-flies condition. There is a discrepancy between the reported sample size (N=106) and the reported degrees of freedom (df=173), which makes it difficult to create simulated experiments to estimate the joint power. As an estimated upper limit, the power of the interaction is reported in Table S28. The power value would not change very much (third decimal only) if the sample size was actually N=177.

Study 4a explored whether belief in the effects of time distortion should influence the reported effects. The study combined data sets from two universities that also differed in whether they were paper-based or computer-based. A key finding was a significant regression between the effect of time manipulation and belief in time distortions. The statistic in Table S28 is a *t* value. Another analysis directly compared enjoyment across time-flies and time-drags conditions. It is not possible to estimate the joint power of these analyses. The power in Table S28 is for the regression only, and it thereby provides an upper estimate of the joint power.

Study 4b also investigated how the time distortion effect on enjoyment was influenced by belief. The key comparison was an interaction between belief in the effect (manipulated by reading a story that either supported or refuted the idea) and the flies/drags conditions. The statistic reported in Table S28 is an *F* value.

Study 5 provided some participants an alternative explanation for why time might seem to be distorted (ear-plugs rather than enjoyment). The time-distortion effect on enjoyment ratings was moderated by the alternative explanation. Table S28 reports the *F* value for an interaction. There were several additional analyses and another measure, so Table S28 provides an upper estimate of their joint power.

Finally, the general discussion section of Sackett *et al.* (2010) reported on additional experiment that was a variation of Study 3, but replaced the decelerated condition with a normal time condition. Table S28 shows the reported *F* statistic.

Table S28: Statistical properties of the Sackett *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1a	18, 19	2.36	.024	0.759	.612
Study 1b	40, 39	2.95	.004	0.657	.822
Study 2	49, 50	5.67	.019	0.475	.648
Study 3	26, 27, 26, 27	4.50	.036	0.409	.550
Study 4a	109	2.57	.012	0.244	.715
Study 4b	15, 15, 15, 15	5.94	.018	0.602	.659
Study 5	15, 15, 15, 15	4.10	.048	0.516	.502
Extra	29, 30	7.67	.008	0.712	.766
P_{TES}					.033

Sackett *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .033$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

28. Analysis of Savani, Markus, Naidu, Kumar and Berlia (2010) "What Counts as a Choice?"

Savani *et al.* (2010) reported six experiments purporting to show that Americans were more likely than Indians to construe actions as choices. Table S29 summarizes the statistical values that contributed to the TES analysis.

Study 1 asked American and Indian participants to list choices they had made the previous day or to list things they had done without any choice. Americans listed more actions in the choice condition than Indians, but were equivalent in the nonchoice condition. The analysis included an interaction and two *t* tests. Power was estimated from 100,000 simulated experiments that used the means and standard deviations provided in the reported data. Study 2 analyzed data with a Mann-Whitney test, and it is not clear how to estimate power without access to the raw data.

Study 3 had a similar design as Study 1, but had participants report action choices or object-touching (control condition) in a video. The analysis included an interaction and two *t* tests (the reported degrees of freedom appear to be incorrect for all of the tests). Success probabilty was estimated from 100,000 simulated experiments that used the means and standard deviations provided in the reported data.

Study 4a used a hierarchical linear model to demonstrate a statistically significant effect that Indians were less likely than Americans to construe actions as choices. The statistic in Table S29 is a *t* value, and the *p* value is a bit over the traditional criterion for significance. Power was estimated with a significance criterion of .053. The regression included tests for other influences as well, so the power value reported in Table S29 sets an

upper limit on the estimated joint power. Study 4b replicated the findings in Study 4a with a different measure. Study 5 reported a significant interaction between culture and conditions of personal- or interpersonal-choice. The statistic reported in Table S29 is a *t* value. For each of these studies it is not clear how the reported degrees of freedom were computed in Savani *et al.* (2010), but it is not going to make much difference to the power calculations.

Table S29: Statistical properties of the Savani *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	40, 39, 50, 50	Multiple tests			.588
Study 3	30, 30,34, 34	Multiple tests			.680
Study 4a	39, 84	1.96	.052	0.377	.499
Study 4b	45, 58	2.20	.030	0.434	.580
Study 5	90, 128	1.98	.049	0.271	.502
P_{TES}					.058

Savani *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .058$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

29. Analysis of Senay, Albarracín, and Noguchi (2010) "Interrogative Self-Talk and Intention"

Senay *et al.* (2010) reported four experiments purporting to show that declarative forms of introspective talk result in goal-directed behavior. Table S30 summarizes the statistical values that contributed to the TES analysis.

Experiment 1 had participants work on anagram puzzles. Participants who prepared for the puzzles by thinking about *whether* they would work on the anagrams solved more problems than participants who prepared by thinking *that* they would work on the anagrams. The statistic in Table S30 is a *t* value.

Experiment 2 had participants solve anagrams after a handwriting task designed to introduce various types of introspective talk. Different groups of participants wrote: *Will I, I will, I,* or *Will*. The analysis involved several hypothesis tests. An ANOVA showed a significant difference in number of anagram solutions across conditions. Follow-up contrasts showed that performance in the *Will I* condition was better than the other conditions and that the other conditions were not statistically different from each other. All of these results were important for the theoretical claims. The joint success probability was estimated with 100,000 simulated experiments that drew samples from populations having the mean and standard deviations described in the original samples.

Experiment 3 predicted that the finding reported in Experiment 2 would be weakened if participants first wrote down a random sequence of numbers rather than a patterned sequence of numbers. In a 2 × 2 design, participants wrote either a random or patterned sequence of numbers and then performed the writing task with either the *Will I* or *I will* conditions. Participants were then asked to report how many hours of exercise they planned to perform for the next week. The analysis reported a significant interaction, and its *F* value is reported in Table S30. The analysis also included some contrasts to characterize the interaction, but one contrast is very strong and the other is not clearly required for the theoretical ideas. The power value in Table S30 should only slightly overestimate the joint success probability of all relevant tests.

Experiment 4 showed that the effect of the interrogative form (*I will*) was due to influences on intrinsic motivation. Participants completed either the *I will* or *Will I* writing task and then reported intentions to exercise. In addition, participants completed questionnaires about their intentions that measured intrinsic and extrinsic motivation. The data were analyzed with a multivariate ANOVA, and mediation analysis. Because of the within-subjects design aspect of the analysis it is not possible to quantitatively estimate the joint success probability of the many tests with the provided summary statistics. Many of the *p* values were just below the criterion for statistical significance (.048, .049, .043). Typically, a *p* value just at the significance criterion indicates a power of .5, and this establishes an upper limit on the joint success probability of the set of relevant tests.

Table S30: Statistical properties of the Senay *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	26, 27	2.11	.040	0.571	.532
Exp. 2	12, 13, 12, 13	Multiple tests			.679
Exp. 3	11, 12, 11, 12	4.14	.048	.590	.498
Exp. 4	28, 28	Multiple tests			.500
P_{TES}					.090

Senay *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .090. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

30. Analysis of West, Anderson, Bedwell and Pratt (2010) "Red Diffuse Light Suppresses Fear Prioritization"

West *et al.* (2010) reported four experiments purporting to show that red diffuse light suppressed rapid perception of fear. Table S31 summarizes the statistical values that contributed to the TES analysis.

In Experiment 1 participants reported the temporal order of two Gabor patches. Previous work reported that Gabors with a low spatial frequency tended to be reported as preceding Gabors with a high spatial frequency that physically co-occurred. This pattern was found when the Gabors were presented on a green background, but disappeared when the Gabors were presented on a red background. The analysis involved three tests that compared the perceived temporal offset against zero for green and red light and also compared the offsets for red and green light against each other. All of the results were statistically significant. The experiment was a within-subjects design, so the measures for red and green backgrounds are correlated. This correlation can be estimated if the standard deviation is assumed to be the same for the red and green backgrounds (r=0.396). The success probability listed in Table S31 is the power for all three tests, and was estimated from 100,000 simulated experiments that drew samples from populations having means, standard deviations, and correlations indicated in the original samples.

Experiments 2a and 2b had a similar design and analysis, but the stimuli were fearful faces and either scrambled fearful (2a) or neutral (2b) faces. The correlation between scores for green and red light were r=0.607, and r=0.421. The power was again estimated with 100,000 simulated experiments.

Experiment 3 investigated red and green background effects on arrow cue priming. As predicted, the cue had an almost equal effect for both background colors. The summary statistics do not include the t value comparing the red and green backgrounds, which was used previously to estimate the correlation between scores for the red and green backgrounds. The t value can be estimated from the given means and using the pooled standard deviation in difference scores for the other three studies (17.12). This gives t=0.26 and r=0.573. With these values and the reported means and standard deviations, 100,000 simulated experiments estimated the joint power of these outcomes. Although there is some uncertainty in this power estimation, different correlation values are unlikely to make much difference in the final conclusion.

Table S31: Statistical properties of the West *et al.* (2010) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	17	Multiple tests			.550
Exp. 2a	15	Multiple tests			.705
Exp. 2b	15	Multiple tests			.476
Exp. 4	10	Multiple tests			.849
P_{TES}					.157

West *et al.* (2010) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .157. Since this value is larger than the .1 criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

31. Analysis of Alter and Oppenheimer (2009) "Fluency and Self-Disclosure"

Alter and Oppenheimer (2009) reported five experiments purporting to show that people self-disclosed less information under conditions of cognitive disfluency than cognitive fluency. Table S32 summarizes the statistical values that contributed to the TES analysis.

In Study 1a participants completed a social-desirability scale that was printed either in a clear font or a difficult-to-read font. The dependent variable was the number of responses to the items in the questionnaire. It is somewhat unusual, but the analysis computed the difference in the percentage of participants in each group who provided an answer for each question and then ran a one-sample *t* test that compared whether the differences across questions were statistically different from zero. It would be more standard to measure the number of responses for each student and then run a two-sample *t* test to look for a difference across conditions. Such an analysis would not have been significant. If the original analysis is taken at face value, then Table S32 summarizes the *t*-test and the *n* term corresponds to the number of items (which happens to equal the number of participants). The analysis also included a measure of frustration, which was (as predicted) not significantly related to the other score. This latter test is not included in the power estimate.

Study 1b had a similar design and analysis but used a different questionnaire. The experiment included thirty-six participants, but the analysis again looked at difference scores across items. A more standard analysis that looked at the average difference across participants in each condition would not have been statistically significant. The statistic in Table S32 is a *t* value for the difference across items analysis, with *n* indicating the number of items on the questionnaire.

Study 2 used the same font manipulation as a between-subjects factor and measured word completion for risk-related words and self-presentational words as a within-subjects factor. The main finding was a significant interaction, and Table S32 reports the *F* value. The analysis included some additional comparisons, but the mixed nature of the design would require knowing the correlations between measures to estimate joint power. Table S32 lists an upper limit of the joint power.

Study 3 used the font manipulation with a different self-disclosure questionnaire. Again, the analysis examined difference scores across items rather than differences between scores in different conditions (the latter analysis would not be statistically significant with the 74 participants). Table S32 lists the *t* value for the difference scores across items, with *n* indicating the number of items on the questionnaire.

Study 4 randomly selected 426 confessions from an on-line confession site that changed its design into an easier to read format. The samples were taken during the two weeks before and after the design change. One hundred participants then rated the confessions for embarrassment (each confession was rated by approximately 10 participants). The analysis compared the embarrassment ratings across the confessions submitted before and after the site redesign. The statistic in Table S32 is an *F* value.

Table S32: Statistical properties of the Alter and Oppenheimer (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1a	33	2.26	.031	0.384	.571
Study 1b	10	2.57	.030	0.743	.554
Study 2	33, 34	4.19	.045	0.494	.513
Study 3	30	2.82	.009	0.501	.756
Study 4	213, 213	4.65	.032	0.209	.574
P_{TES}					.071

Alter and Oppenheimer (2009) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .071$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

32. Analysis of Ashton-James, Maddux, Galinsky and Chartrand (2009) "Affect and Culture"

Ashton-James *et al.* (2009) reported four experiments purporting to show that positive affect promoted thoughts and behavior that departed from cultural constraints but negative affect bound people to cultural norms. Table S33 summarizes the statistical values that contributed to the TES analysis.

Experiment 1 assigned Western or Eastern culture participants to one of three affect conditions (positive, neutral, or negative) and had them complete a questionnaire that measured the value of self-expression. The analysis separately investigated the Western and Eastern culture participants. For the Western participants, significant differences were expected and found for all comparisons across conditions. Power was estimated with 100,000 simulated experiments that drew samples from populations having means and standard deviations matching the original samples. The power for the positive vs. negative comparison is quite high, but the power for each of the neutral comparisons is around .6. The joint power for all of the comparisons to be significant is quite low because the scores for the neutral condition have to be placed nicely between the scores for the positive and negative conditions in order to be statistically different from both of them. It should be noted that the reported degrees of freedom for the tests in Ashton-James *et al.* (2009) do not match the reported sample sizes.

Table S33 also reports the *F* value for Eastern participants in the negative and positive affect conditions (the effect is the opposite of the Western participants). Comparisons with the neutral condition were not statistically significant, but this apparently was not required by the theory. The test statistic in Table S33 is an *F* value.

Experiment 2 varied culture (Western and Eastern) and affect (negative and positive) and measured how many participants selected an uncommon colored pen from a set. The analysis reported a significant difference in proportions across affect conditions for

Westerners. Table S33 reports a χ^2 value, and the listed effect size is Cohen's *h* (n_1=15, p_1=0.600, n_2=15, p_2=0.133). For the theory, it was also important that the Eastern culture participants did not show a significant result in the same direction, and the probability of this outcome is almost one because the (non-significant) trend was in the opposite direction.

Experiment 3 varied affect by having Western and Eastern participants make a frowning-type or smiling-type face. Participants then generated responses that were scored by a measure of self-construal. The dependent variable was the proportion of responses that indicated independent (as opposed to interdependent) self-construal. The analysis reported a significant interaction between culture and affect, with Westerners showing significantly fewer independent self-construals for positive affect than for negative affect and Easterners showing a significant opposite relation. Although the interaction and Easterners effects were quite strong, the *p* value for the Westerners effect was not much below the .05 criterion, so the power of the joint outcome is rather modest. Power was estimated with 100,000 simulated experiments that drew samples from populations having the mean and standard deviations of the original samples.

Experiment 4 varied affect by having participants view positive or negative images. The distance a participant sat next to a confederate on a bench constituted a measure of self-construal. An ANOVA showed a significant interaction between culture and affect, and contrasts showed significant opposite directions for Westerners and Easterners. The latter comparison was judged significant with a non-standard criterion of .06, and this value was used in the power calculations. Power was estimated with 100,000 simulated experiments that drew samples from populations having the mean and standard deviations of the original samples. Although the interaction was highly significant, each of the effects within culture were modest (but deemed necessary for support of the theory), thus, the power is quite low.

Table S33: Statistical properties of the Ashton-James *et al.* (2009) experiments.

	n	**Test Statistic**	**p**	**Effect size**	**Probability of Success**
Exp. 1 Westerners	23, 24, 25	Multiple tests			.290
Exp. 1 Easterners	27, 28	12.36	<.001	0.935	.925
Exp. 2 Westerners	15, 15	7.033	.008	1.02	.801
Exp. 2 Easterners	15, 18	1.782	.18	-0.483	1.00
Exp. 3	30, 30 15, 16	Multiple tests			.495
Exp. 4	8, 9, 9, 9	Multiple tests			.331
P_{TES}					.035

Ashton-James *et al.* (2009) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .035. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

33. Analysis of Fast and Chen (2009) "Power, Incompetence, and Aggression"

Fast and Chen (2009) reported four experiments purporting to show that people become aggressive if they feel incompetent in a domain where they have power. Table S34 summarizes the statistical values that contributed to the TES analysis.

Study 1 surveyed working people about power, aggression, and self-perceived competence. A regression analysis identified several significant results that are related to the theoretical ideas. The finding that power was associated with aggression among people with high self-perceived incompetence sets an upper limit on statistical power for the findings. Table S34 reports the *t* value supporting this hypothesis test.

Study 2 primed participants for power (high or low) and competence (high or low) with a structured writing assignment. As a measure of aggression, participants then chose a horn blast that, ostensibly, would be used to indicate errors in a future experiment. Louder sounds indicated higher aggression. An ANOVA reported a significant interaction and four independent *t*-tests that contrasted the power and competence conditions. (The given degrees of freedom suggest that these *t*-tests were not contrasts, which would have used the *MSE* from the ANOVA). Consistent with the theory, two of these *t*-tests produced significant results and two of the tests produced non-significant results. The joint success probability for all of these tests was computed from 100,000 simulated experiments that sampled from populations having the means and standard deviations of the original samples. For one condition (low-power, high competence) the mean was taken from Figure 2 (Fast & Chen, 2009) and the standard deviation (0.77) was derived from the means and the *t* value comparing the high competence conditions that differed in power. The low success probability in Table S34 is due to several comparisons producing findings in the original sample that just barely meet the criterion for statistical significance. For them all to be significant is uncommon for most samples.

Study 3 placed participants in a leadership position with power but then provided them with either positive or negative feedback about leadership aptitude. Measures included self-perceived competence and whether the participant selected at least one difficult task for another student (a proxy for aggression). A logistic regression reported a significant interaction between competence and feedback. The statistic in Table S34 is a Wald test value. It is not clear how to estimate power for a Wald test, but since the *p* value for the test is right at the criterion threshold, a good estimate is that power is .5. The analysis included several other tests as well, so this estimate establishes an upper bound on the joint power.

In Study 4 participants rated their power at work, were treated to conditions that varied in primed competence and self-worth, and then completed a survey on aggression. The key result was a three-way interaction, and Table S34 includes the *t* value for the test. The analysis examined other effects as well, so the statistical power value in Table S34 is an upper estimate on the joint power.

Fast and Chen (2009) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .072$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

Table S34: Statistical properties of the Fast and Chen (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	90	2.48	.015	0.259	.682
Study 2	22, 25, 24, 27	Multiple tests			.328
Study 3	32, 27	3.76	.050	--	.500
Study 4	163	2.36	.019	0.184	.646
P_{TES}					.072

34. Analysis of Fast, Gruenfeld, Sivanathan and Galinsky (2009) "Power and Illusory Control"

Fast *et al.* (2009) reported four experiments purporting to show that the experience of power produces an illusion of personal control. Table S35 summarizes the statistical values that contributed to the TES analysis.

Experiment 1 assigned participants to a high-power, low-power, or baseline priming condition and then asked them to either roll a six sided die (to possibly win a prize) themselves or to have the experimenter roll the die. The decision served as a measure of illusory control. The analysis compared the proportion of participants across conditions that chose to roll the die themselves. A significant difference was reported across conditions, with additional tests revealing that high-power participants were more likely to roll the die themselves than participants in the low-power or baseline conditions (the reported statistics for the latter comparison appear to be incorrect, both the χ^2 and p value should be smaller). The low-power and baseline proportions were not significantly different, but it is not clear that any outcome of that test is meaningful for the theory. The success probability of the first three tests was estimated with 100,000 simulated experiments that drew samples from populations having the proportions indicated in the original samples.

Experiment 2 manipulated power by assigning participants to a role as a manger (had power) or a worker (without power). Participants then read a (ostensibly unrelated) vignette involving a marketing agency and provided ratings about control and optimism relative to their (unspecified) role in the agency. Table S35 provides the t value comparing the optimism measure across the groups. The analysis included additional tests related to control and mediation effects, but it is not possible to compute the joint success probability of all the effects without knowing the correlation between variables. The value listed in Table S35 provides an upper limit on the estimated joint success probability.

Experiment 3 manipulated power and then measured sense of control, self-esteem, and action orientation. All three dependent variables showed significant differences between the high- and low-power conditions. It is not possible to estimate the joint power for all three tests without knowing the correlation between the variables. An upper limit of the success probability can be estimated by looking at the test with the lowest success probability, which is for the action orientation measure. The test statistic in Table S35 is a χ^2 value, and the listed effect size is Cohen's h (n_1=40, p_1=0.775, n_2=39 p_2=0.564).

Experiment 4 assigned participants to a high-power or baseline context and had them answer a questionnaire that measured optimism, self-esteem, and mood. The first two

measures showed statistical significance but the third did not, which was consistent with the theory. The joint success probability of all measures cannot be estimated, but an upper limit can be estimated by the success probability of the self-esteem test. Table S35 lists the *t* value for the test.

Table S35: Statistical properties of the Fast *et al.* (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	13, 12, 13	Multiple tests			.440
Exp. 2	15, 15	2.29	.029	0.814	.576
Exp. 3	40, 39	3.98	.046	0.454	.523
Exp. 4	23, 20	2.11	.041	0.633	.525
P_{TES}					.069

Fast *et al.* (2009) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .069$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

35. Analysis of Garcia and Tor (2009) "The N-Effect"

Garcia and Tor (2009) reported seven experiments purporting to show that competitive motivation decreased as the number of competitors increased. Table S36 summarizes the statistical values that contributed to the TES analysis.

Studies 1a and 1b computed correlations between the estimated number of test takers and scores on standardized tests. The analyses controlled for several other variables (e.g., gender, state funding, income). The statistic in Table S36 is an *r* value that was computed after factoring out the control variables. The sample size in Study 1a refers to a score from each of the fifty U.S. states while the sample size in Study 1b refers to a score from each of twenty-two test taking opportunities. The significance tests of the correlation values used reduced degrees of freedom to compensate for the contribution of the control variables, and this somewhat complicates the power estimation. The power values in Table S36 did not consider the reduced degrees of freedom, which tends to maximize the power values. Simulations suggested that this approximation makes little difference to the power value.

Study 2 asked participants to quickly complete a quiz, where the top 20% would receive a prize. Across conditions, participants were told that they were competing against either ten or one hundred other participants. Participants were faster in the 10-person condition, and the statistic in Table S36 is an *F* value.

A footnote to Study 2 described another experiment where participants read about Study 2 and then rated how concerned they would be about their performance compared to

any other competitor. These social-comparison concerns were higher in the 10-person condition. The statistic in Table S36 is a *t* value.

Study 3 had participants read vignettes that described a competition against 50 or 500 others and asked participants to estimate how hard they would compete. Each participant provided a rating for the large and small competitor group and also provided a social-comparison orientation scale value, which was used as a covariate in the analysis. After factoring out the social-comparison orientation value, there was a significant difference in the conditions. The statistic in Table S36 is an *F* value.

Study 4 used two different samples of participants who were asked social comparison questions about competitors for a job interview. The main effect of number of competitors was very strong for both samples. A key comparison for each sample reported a null result for an effect of behavior on a ratio-bias scenario (a possible alternative explanation of the reported effect). These tests produced *p* values that were not significant (.20 and .16). Unfortunately, the provided statistics do not allow for a success probability estimation because the text does not include an effect size measure and the effect size cannot be computed from the provided summary statistics. The probability of the null results is surely less than one, so the TES analysis will overestimate the P_{TES} value.

Study 5 used a similar competition story as the earlier studies and asked participants to rate several variables: motivation to compete, compare progress to competitors, and ease of winning the competition. The analyses reported significant differences for all of these variables. It is not possible to estimate the joint success probability for all three effects, so Table S36 shows the *F* value for the weakest effect, which gives an upper estimate on the joint success probability.

Table S36: Statistical properties of the Garcia and Tor (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1a	50	-.35	.014	--	.708
Study 1b	22	-.48	.024	--	.625
Study 2	37, 37	4.09	.047	0.465	.506
Footnote	30, 30	3.44	.001	0.877	.916
Study 3	47	5.41	.025	0.334	.610
Study 5	25, 25	6.82	.012	0.727	.712
P_{TES}					.089

Garcia and Tor (2009) reported that each of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .089. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

36. Analysis of Gonzáles and McLennan (2009) "Hemispheric Differences in Sound Recognition"

Gonzáles and McLennan (2009) reported four experiments purporting to show hemispheric asymmetries for nonlinguistic auditory stimuli. Table S37 summarizes the statistical values that contributed to the TES analysis. Every experiment made a similar prediction for the pattern of results across several hypothesis tests: a significant difference between two priming conditions presented to the right hemisphere (left ear), but no difference when presented to the left hemisphere (right ear). Gonzáles and McLennan (2009) ran two types of analyses for each experiment. One analysis used participants as the sampled variable and the other analysis used test items as the sampled variable. With only one exception, the analyses produced almost identical results, and the TES analysis is based on the item analyses only. Table S37 reports the within-item *t* values and post hoc power for each experiment. The listed *n* is the number of test items. The reported degrees of freedom in Gonzáles and McLennan (2009) are sometimes incorrect for the item analyses.

Experiment 4 is the exception noted above. Although it produced an effect in the predicted direction, it was not statistically significant across participants but was marginally significant across items. The text discussing the results indicates that Gonzáles and McLennan (2009) feel the marginally significant result across items provides positive support their theoretical ideas, and they do not provide a quantitative model (such as a random effects model) that would accommodate a null result. Instead arguments are made about variations in the test and task conditions that differ from the other experiments. For the TES analysis described in Table S37, the marginally significant result for Experiment 4 is considered to meet the criterion for significance ($p<.1$).

In addition, each experiment had two predicted null findings. Given the within-subjects design of the experiments, it is not possible to compute the post hoc probability of a non-significant finding, but if the null really is true, then each experiment has a 95% chance of not rejecting the null. The penultimate row in Table S37 lists the probability that all eight tests should produce non-significant findings. The probability of the entire set of results is the product of these values, as listed in the final row of Table S37.

Table S37: Statistical properties of the Gonzáles and McLennan (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	24	3.56	.002	0.703	.909
Exp. 2	24	2.74	.011	0.541	.718
Exp. 3	24	2.29	.029	0.407	.615
Exp. 4	24	1.81	.084	0.357	.522
Predicted nulls		Eight tests		0.000	.663
P_{TES}					.139

Gonzáles and McLennan (2009) reported that all of the studies produced results supporting their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to produce the

desired pattern is the product of the probabilities: P_{TES} = .139. Since this value is larger than the .1 criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

An alternative analysis could consider the tests across participants, in which case the finding in Experiment 4 should be considered non-significant. Such a non-significant finding is often considered a failure, but for the TES analysis it reflects an expected outcome of relatively low powered studies. The probability that three or more experiments would reject the null hypothesis from a set of four experiments having power values like these can be computed with a Fisher's exact test that compares all four possible combinations of three out of four experiments and one additional combination where all four experiments reject the null. The sum of these probabilities is .48. When combined with the null results, the TES analysis concludes P_{TES} = .319.

37. Analysis of Hahn, Close and Graf (2009) "Transformation Direction"

Hahn *et al.* (2009) reported four experiments purporting to show that perceived similarity of two images is influenced by the direction of a linking transformation. Table S38 summarizes the statistical values that contributed to the TES analysis.

In Experiment 1 participants watched a movie where a target object morphed into another object or vice-versa. Participants then made a similarity judgment about the objects. A significant difference was reported for the different morphing directions. The statistic in Table S38 is an *F* value. Experiment 2 was a near replication of Experiment 1 that ruled out a possible methodological confound. Experiment 3 was also a close replication, with a control to rule out concerns about spatial overlap.

Experiment 4 was designed to rule out alternative explanations of the effect by randomly scrambling some of the movie frames and thereby breaking the perceived transformation of the morphing. The data supported the theory with a null finding. Table S38 reports the *F* statistic, and the reported success probability value is the probability of not rejecting the null.

Table S38 : Statistical properties of the Hahn *et al.* (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	20, 20	6.89	.012	0.814	.708
Exp. 2	24, 24	10.26	.002	0.910	.870
Exp. 3	24, 24	8.11	.007	0.809	.783
Exp. 4	24, 24	2.02	.162	0.404	.722
P_{TES}					.348

Hahn *et al.* (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .348. Since this value is larger than the .1

criterion, the TES analysis does not indicate that readers should be skeptical about the reported experimental results as they relate to the theory.

38. Analysis of Hart and Albarracín (2009) "Describing Actions"

Hart and Albarracín (2009) reported four experiments purporting to show that using different verbal forms promote memory for action-relevant knowledge. Table S39 summarizes the statistical values that contributed to the TES analysis.

In Experiment 1 participants wrote an essay about a non-stereotypical African American male and then were asked to describe their previous behavior in terms of what they *did* (perfective aspect) or what they *were doing* (imperfective aspect, which implies it will continue). Finally, participants rated hostility for an African American character in the "Donald" story, where lower hostility ratings implied reduced activation of the African American stereotype. The results found lower hostility ratings in the perfective condition than in the imperfective condition, which implies that the perfective form promotes carry-over across tasks (avoiding stereotyping). The statistic in Table S39 is an F value.

In Experiment 2 participants working on an anagram task were interrupted and asked to describe their actions with either the imperfective or perfective aspect. The dependent variable was whether participants were then willing to go back to the anagram task. A χ^2 test found a significant difference in the proportion of participants willing to continue. In Table S39 the effect size is Cohen's h (n_1=20, p_1=0.700, n_2=17, p_2=0.353). This experiment also reported a significant difference in memory for the anagrams, but the joint power cannot be estimated without knowing the correlation between the measured variables. The power in Table S39 is an upper limit of the joint power.

Experiment 3 was similar to Experiment 2 but measured performance on a lexical decision task. As predicted, reaction times for words were shorter for participants who described their previous work with the imperfective rather than the perfective. The statistic in Table S39 is an F value, which is the result of an ANCOVA. The effect size is standardized relative to the standard deviation of the adjusted scores, which means it cannot be compared to other experiments (which would use other standardizations) but can be used to estimate power. This experiment also included many other tests, but the joint success probability cannot be calculated without knowing the correlations between variables.

Experiment 4 used a 2 × 2 design that varied the relevance of information to the anagram task and the aspect of language (imperfective vs. perfective) used to describe that information. The dependent variable was performance on a subsequent anagram task. The analysis included a significant interaction between aspect and relevance and four significant contrasts (comparing all rows and columns of the 2 × 2 design). For two of these contrasts significance was concluded for a criterion of p<.11, and this criterion was used for the success probability calculations. Success probability was estimated with 100,000 simulated experiments that sampled data from populations having the means and standard deviations described in the original samples. Each contrast by itself had power just above one half, but the success probability for the entire set is quite low because many restrictions are being placed on the results in order for them to be consistent with the theory.

Table S39: Statistical properties of the Hart and Albarracín (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	25, 26	4.30	.043	0.572	.517
Exp. 2	20, 17	4.46	.035	0.710	.576
Exp. 3	74, 74	4.87	.029	0.365	.587
Exp. 4	40, 40, 40, 39	Multiple tests			.201
P_{TES}					.035

Hart and Albarracín (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .035$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

39. Analysis of Janssen and Caramazza (2009) "Phonology and Grammatical Encoding"

Janssen and Caramazza (2009) reported four experiments purporting to show that grammatical encoding is sensitive to the phonological properties of words. Table S40 summarizes the statistical values that contributed to the TES analysis.

In Experiment 1a participants viewed an object and a word. The task for one group of participants was to quickly name the object and then read the word. Another group of participants viewed a colored word and first named the color and then read the word. In addition, the picture and word, or color and word, were either related (shared a common onset consonant) or unrelated. The dependent variable was reaction time; errors were also measured but seemed to play a lesser role in the theoretical claims. As is common in linguistic-related studies, the analysis for all of the experiments reported tests for both subjects and items. The TES analysis assumed that significant findings were required in both types of analyses unless the text indicated otherwise, and indeed this is generally what was reported. The analysis across items identified a significant interaction, and subsequent *t*-tests found a significant difference in relatedness for the object condition. Another significant difference was found for the color condition, but in the opposite direction. An upper limit for power can be estimated with the weakest outcome, which is the significant difference reported for noun-noun utterances. The statistic in Table S36 is a *t* value.

Experiment 1b had a similar task, but varied the content of the pictures to investigate noun-verb interactions (whereas Experiment 1a investigated noun-noun and noun-adjective interactions). The analysis found a significant effect of consonant relatedness. The statistic in Table S40 is a *t* value across subjects, and the *n* value is the number of participants. A similar result was found across items, so the estimated power slightly overestimates the joint power.

Experiment 2a compared reaction times for adjective-noun versus noun-adjective utterances. The analysis across items reported a marginally significant interaction, and

follow-up *t* tests found a significant difference between relatedness conditions for the adjective-noun condition but not for the noun-adjective condition. Not enough information is provided to compute the success probability for all of these tests. The success probability in Table S40 is for the interaction only. Since the test for the interaction was described as "marginally" significant, the power estimation used a criterion of $p<.075$.

Experiment 2b had participants produce adjective-adjective-noun utterances. The analysis identified a main effect for relatedness. The statistic in Table S40 is a *t* value for the item analysis. Again, significance was determined with a non-standard criterion. The power estimation used a criterion of .075.

Table S40: Statistical properties of the Janssen and Caramazza (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1a	40, 40	1.97	.052	0.436	.497
Exp. 1b	15	2.60	.021	0.635	.628
Exp. 2a	80, 80	3.50	.063	0.294	.529
Exp. 2b	40, 40	1.82	.073	0.403	.502
P_{TES}					.083

Janssen and Caramazza (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .083$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

40. Analysis of Jostmann, Lakens and Schubert (2009) "Weight and Importance"

Jostmann *et al.* (2009) reported four experiments purporting to show that the abstract concept of importance is grounded in physical experiences of weight. Table S41 summarizes the statistical values that contributed to the TES analysis.

Every experiment had participants hold either a heavy or light clipboard and answer questions on the importance of various topics. Given the common task and design aspects, it seems appropriate to pool the effect sizes across Studies 1, 2, and 4 using a meta-analytic technique that gives 0.646 as the best estimate of the effect size. The effect sizes of the reported experiments never deviate much from this value. Although it gives a very similar effect size, study 3 used a different analysis, so the effect size should not be pooled.

Studies 1, 2, and 4 all reported simple comparisons between clipboard conditions. Table S41 reports *F* statistics. Study 4 also included several other measures and hypothesis tests, so the reported power is an upper limit estimate of the joint power.

Study 3 reported an interaction from a regression analysis, which found a significant effect of clipboard weight. The statistic listed in Table S41 is a *t* value.

Jostmann *et al.* (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as

estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .090. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

Table S41: Statistical properties of the Jostmann *et al.* (2009) experiments. Power is estimated relative to the pooled effect size for studies 1, 2, and 4. Post hoc power is used for Study 3.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	20, 20	4.86	.034	0.683	.512
Study 2	28, 23	4.33	.043	0.577	.614
Study 3	25, 24	2.19	.034	0.616	.560
Study 4	20, 20	5.06	.030	0.697	.512
P_{TES}					.090

The conclusion of bias in this article is confirmed by the first author's report of a study that did not reject the null hypothesis but which was not discussed in the paper. A description of the non-significant experiment is posted on the PsychFiledrawer web site (Jostmann, 2013). As the author noted, "At hindsight we should have mentioned this study in the original submission. Back then, I (admittedly) naively concluded that the attempt was not worth mentioning because I blamed the noisy conditions at the train station. I certainly know better now." An interesting discussion of this article and replication attempts can be found in Zwann (2013).

41. Analysis of Labroo, Lambotte and Zhang (2009) "The Name-Ease Effect and Importance Judgments"

Labroo *et al.* (2009) reported six experiments purporting to show that providing a name for a result produces feelings of ease. Table S42 summarizes the statistical values that contributed to the TES analysis.

Experiment 1a divided participants in a 2 × 2 design that varied whether a court case was named or not and whether the participants took part in a memorability or an understandability task. The dependent variable was reported importance of the case. The analysis showed a significant interaction, and contrasts showed increased importance for participants in the memorability condition but an opposite effect for the understandability condition. All of these results were significant. The joint success probability of the findings was estimated with 100,000 simulated experiments that drew samples from populations having the means and standard deviations of the original samples. Just as in the original analysis, the contrasts were evaluated for significance with one-tailed *t* tests. The joint success probability is low because the *t* tests each produced a *p* value just below the standard criterion.

Experiments 1b and 1c replicated the findings of Experiment 1a with new participants and different named or unnamed findings. Experiment 2 also had a very similar

design and analysis. Success probability was estimated in the same way. Experiment 2 also included a measure of ease of memorability or understanding and a mediation analysis. It is not possible to estimate the joint success probability of the principle effects and the mediation, so the power in Table S42 provides an upper estimate of the success probability for Experiment 2.

Experiment 3a had a similar design, but also measured reported attention to the experimental task. A regression analysis reported a significant interaction between naming and task, and the statistic in Table S42 is a *t* value. Significance was judged with a one-tailed test. The degrees of freedom of the test depend on the number of other variables in the regression. It is not entirely clear how sample size contributes to the effect size for this case. The effect size in Table S42 uses the given sample size to compute the effect size and power. It turns out that other choices would make little difference. The original analysis also included several other significant findings, but it is not possible to estimate the joint power of all the effects. The power value given in Table S42 should be considered an upper estimate.

Experiment 3b was similar to the other experiments but included another factor that manipulated attention. The analysis reported a significant interaction of task and name conditions and a three-way interaction of those variables with attention. The pooled standard deviation (1.33) was estimated from the reported *t* tests. For the high attention condition, the (one-tailed) contrasts showed the expected naming effects for the different tasks. For the low attention condition, the theory predicted no effect for the contrasts, and this is what was reported. The probability of not producing a significant result is relatively high for each contrast, but their joint probability is substantially reduced. The success probability in Table S42 is the estimated probability (from 100,000 simulated experiments) of getting the full range of significant and non-significant outcomes that Labroo *et al.* (2009) reported were relevant for their theoretical conclusion.

Table S42: Statistical properties of the Labroo *et al.* (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1a	10, 10, 9, 8	Multiple tests			.328
Exp. 1b	8, 8, 9, 9	Multiple tests			.412
Exp. 1c	10, 10, 8, 7	Multiple tests			.334
Exp. 2	21, 20, 21, 17	Multiple tests			.745
Exp. 3a	84	1.77	.040	.191	.538
Exp. 3b	9, 9, 9, 9 10, 9, 9, 9	Multiple tests			.410
P_{TES}					.008

Labroo *et al.* (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .008$. Since this value is smaller than the .1

criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

42. Analysis of Nordgren, van Harreveld and van der Pligt (2009) "Restraint Bias"
Nordgren *et al.* (2009) reported four experiments purporting to show that impulse-control beliefs influence processes of self-control. Table S43 summarizes the statistical values that contributed to the TES analysis.

Study 1 assigned participants to a manipulation that induced mental fatigue or not. Participants were then asked what percentage of time for studying they intended to leave to the end of the semester. Fatigue participants reported a smaller percentage than non-fatigued participants, thereby indicating less perceived control. The statistic in Table S43 is an *F* value. Study 1 also included other measures and tests, but it is not possible to estimate their joint power. The power value in Table S43 is an upper limit estimate.

Study 2 asked participants who were either hungry or sated to perform several tasks. First, they completed a questionnaire to measure impulse-control beliefs and hunger cravings. Second, they picked a snack from a set of self-ranked choices. Third, one week later they had the option to return the same snack for cash. Given the correlations between these measures, it is not possible to compute the joint power estimate. As an upper limit of the joint power, the TES analysis focused on the snack return data. The test statistic listed in Table S43 is a χ^2 value, and the listed effect size is Cohen's *h* (n_1=38, p_1=0.605, n_2=41 p_2=0.390). The text in Nordgren *et al.* (2009) is somewhat inconsistent about the criterion for statistical significance. The discussion of Study 2 notes that the finding is not significant (p=.06), but the article's conclusion indicates a significant effect and does not refer to the non-significant result. Moreover, Study 3 reports the same *p* value, but it is described as statistically significant. Overall, it seems that Nordgren *et al.* (2009) considered the p=.06 value to be significant. For the TES analysis, the criterion for significance of both Study 2 and 3 was .06.

Study 3 also measured several variables from smokers who were in either a high- or low-control condition. One analysis was the failure to abstain from smoking while watching a film. The test statistic listed in Table S43 is a χ^2 value, and the listed effect size is Cohen's *h* (n_1=27, p_1=0.333, n_2=26 p_2=0.115). It is not possible to estimate the joint power across the other measured variables, so the power value listed in Table S43 should be considered an estimated upper limit.

Study 4 measured impulse-control beliefs and amount of routine smoking temptation among a group of abstaining smokers. These variables were significantly correlated, and the test statistic in Table S43 lists the correlation. The study also measured smoking rates four months later and found additional significant effects. It is not possible to estimate the joint power of these tests, so the power value in Table S43 should be considered as an estimate of the upper limit.

Nordgren *et al.* (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: P_{TES} = .0998. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

Table S43: Statistical properties of the Nordgren *et al.* (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	36, 36	4.51	.037	0.495	.545
Study 2	38, 41	3.65	.056	0.433	.518
Study 3	27, 26	3.59	.058	0.538	.531
Study 4	55	-.32	.017	--	.667
P_{TES}					.0998

43. Analysis of Wakslak and Trope (2009) "Construal Level and Subjective Probability"

Wakslak and Trope (2009) reported five experiments purporting to show that adopting a higher-level construal mind-set made participants generate lower probability assessments. Table S44 summarizes the statistical values that contributed to the TES analysis.

Study 1 asked participants to compare two cameras that were to be used for a raffle. The camera models differed in order to introduce varied levels of abstraction (alignable comparisons introduced a lower-level mind-set while nonalignable comparisons introduced a higher-level mind-set). Participants were then asked how likely it was for a group to sponsor the raffle. As predicted, participants in the nonalignable group reported lower probabilities of raffle sponsorship. The statistic in Table S44 is a *t* value. Since the original analysis notes that the *t* value was based on samples with unequal variances, the power was estimated with 100,000 simulated experiments that drew samples from populations having the means and standard deviations of the original samples.

Study 2 manipulated mind-set by having participants generate superordinate categories for words (high-level construal) or subordinate exemplars for words (low-level construal). Participants then completed a questionnaire to indicate the likelihood that characters would complete certain activities. As predicted, the likelihood judgments were lower for participants in the high-level construal condition. The statistic in Table S44 is a *t* value.

Study 3 manipulated mind-set by asking a series of question that either became more abstract (high-level) or more concrete (low-level). Participants then judged the likelihood of an event happening to a character. As predicted, the likelihood judgments were smaller for the high-level condition. The statistic in Table S44 is a *t* value. The analysis also included some related tests, but they do not seem to be central to the theoretical ideas.

Study 4a manipulated mind-set by having participants identify letters defined globally or locally. Participants then made likelihood judgments for two different situations. The main finding was the predicted effect of mind-set, and Table S44 reports the *F* value. The reported *p* value is described as being below the .05 criterion, but a recalculation gives .05198, and even supposing the reported *F* was rounded down does not allow the *p* value to reach the traditional significance criterion. The power estimation treated the significance criterion as .052. The original analysis included other comparisons

of the ANOVA, but they were not deemed important for the theoretical ideas. Study 4b used the same design, but with a different mind-set inducing task. The statistic in Table S44 is an *F* value.

Table S44: Statistical properties of the Wakslak and Trope (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Study 1	13, 13	2.21	.040	--	.556
Study 2	45, 45	2.62	.010	0.548	.729
Study 3	13, 13	2.44	.022	0.927	.621
Study 4a	11, 11	4.27	.051	0.848	.480
Study 4b	29, 29	4.11	.047	0.525	.502
P_{TES}					.061

Wakslak and Trope (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .061$. Since this value is smaller than the .1 criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

44. Analysis of Zhou, Vohs and Baumeister (2009) "Symbolic Power of Money"

Zhou *et al.* (2009) reported six experiments purporting to show that reminders of money altered social distress and physical pain. Table S45 summarizes the statistical values that contributed to the TES analysis.

Experiment 1 exposed participants to a social rejection or acceptance condition and then measured desire for money in three ways. The three measures were analyzed with separate *t* tests. The correlations between measures were given as a range of values, so 100,000 simulated experiments were created to estimate the probability of all three measures producing significant results for a sample.

In Experiment 2 participants performed a word completion task that primed neutral or pain concepts. Desire for money was measured with two questions, and each question showed significance in separate *t* tests. The original analysis included the correlation between the two measures, so the joint power was estimated with 100,000 simulated experiments.

Experiment 3 manipulated thinking about money by having participants count bills or pieces of paper. Participants then experienced a computer game with or without ostracism. The dependent variables were a rating of social distress, counts of events in the game, and scores of affect. The analysis focused on the distress ratings, and the theoretical ideas were supported by a significant interaction between the counting task and game condition. Planned contrasts further verified that the interaction was due to changes in the ostracism condition across counting tasks and not to the non-ostracism conditions. The pooled standard deviation (0.628) was estimated from a reported contrast *F* value and the

provided means for the conditions. These values contributed to the success probability estimation of the three results with 100,000 simulated experiments. The relatively low success probability for the tests is mostly due to the test for the interaction, which produced a *p* value not much below the criterion. There were additional tests on the other dependent variables, but they did not run counter to the theoretical ideas and it is not possible to compute their joint success probability.

Experiment 4 used the same counting tasks and then introduced participants to either a high-pain or moderate-pain condition. The dependent variables were pain ratings and affect scores. The data analysis was similar to Experiment 3, with a significant interaction and two planned contrasts. The pooled standard deviation (1.178) was estimated from a reported contrast *F* value and the sample means. The joint success probability was estimated from 100,000 simulated experiments. Additional data analyses were reported for the affect scores, but they did not run counter to the theoretical ideas and it is not possible to compute the joint success probability.

Experiment 5 was similar to Experiment 3 but varied whether participants wrote about losing money or about weather conditions. One key test was for a difference in the affect scores for the item *strong* across the writing conditions. The statistic in Table S45 is an *F* value. Because of differences between the sample size and degrees of freedom, success probability was estimated from 100,000 simulated experiments using the means and standard deviations of the original samples. There were several tests on other dependent variables as well, so the power value listed in Table S45 should be considered an upper estimate of the joint success probability.

Experiment 6 was similar to Experiment 5, but measured physical pain rather than social distress. One key result was a significant interaction between the pain and writing conditions. Table S45 reports the *F* value. Success probability was estimated from 100,000 simulated samples that used the means and pooled standard deviation (1.273) from the reported data.

Table S45: Statistical properties of the Zhou *et al.* (2009) experiments.

	n	Test Statistic	p	Effect size	Probability of Success
Exp. 1	36, 36	Multiple tests			.424
Exp. 2	46, 46	Multiple tests			.828
Exp. 3	21, 21, 21, 21	Multiple tests			.519
Exp. 4	24, 24, 24, 24	Multiple tests			.592
Exp. 5	54, 54	5.03	.027	--	.609
Exp. 6	24, 24, 24, 24	5.25	.024	--	.619
P_{TES}					.041

Zhou *et al.* (2009) reported that all of the studies produced results that supported their theoretical stance. If the theory is correct, and the population effects are as estimated by the samples, then the probability of getting studies like these to all produce the desired pattern is the product of the probabilities: $P_{TES} = .041$. Since this value is smaller than the .1

criterion, the TES analysis indicates that readers should be skeptical about the reported experimental results as they relate to the theory.

References not in the main article

Arneson, J. J., Sackett, P. R., & Beatty, A. S. (2011). Ability-performance relationships in education and employment settings: Critical tests of the more-is-better and the good-enough hypotheses. *Psychological Science, 22*(10), 1336-1342.

Arnold, D. H., Erskine, H. E., Roseboom, W., & Wallis, T. S. A. (2010). Spatiotemporal rivalry: A perceptual conflict involving illusory moving and static forms. *Psychological Science, 21*(5), 692-699.

Bahrami, B., Vetter, P., Spolaore, E., Pagano, S., Butterworth, B., & Rees, G. (2010). Unconscious numerical priming despite interocular suppression. *Psychological Science, 21*(2), 224-233.

Balcetis, E. & Dunning, D. (2012). A false-positive error in search in selective reporting: A refutation of Francis. *i-Perception,* Downloaded from http://i-perception.perceptionweb.com/journal/I/article/i0519ic.

Benenson, J. F., Markovits, H., Emery Thompson, M., & Wrangham, R. W. (2011). Under threat of social exclusion, females exclude more than males. *Psychological Science, 22*(4), 538-544.

Chambers, J. R. (2010). Why the parts are better (or worse) than the whole: The unique-attributes hypothesis. *Psychological Science, 21*(2), 268-75.

Champely, S. (2009). pwr: Basic functions for power analysis. *R package version 1.1.1*. http://CRAN.R-project.org/package=pwr.

Chen, X., & Hegdé, J. (2012). Learning to break camouflage by learning the background. *Psychological Science, 23*(11), 1395-1403.

Chen, Z., & Treisman, A. (2009). Implicit perception and level of processing in object-substitution masking. *Psychological Science, 20*(5), 560-567.

Endress, A. D., & Potter, M. C. (2012). Early conceptual and linguistic processes operate in independent channels. *Psychological Science, 23*(3), 235-245.

Eyal, T., & Epley, N. (2010). How to seem telepathic: Enabling mind reading by matching construal. *Psychological Science, 21*(5), 700-705.

Finkbeiner, M., & Palermo, R. (2009). The role of spatial attention in nonconscious processing: A comparison of face and nonface stimuli. *Psychological Science, 20*(1), 42-51.

Gajewski, D. A., Philbeck, J. W., Pothier, S., & Chichka, D. (2010). From the most fleeting of glimpses: On the time course for the extraction of distance information. *Psychological Science, 21*(10), 1446-1453.

Ginges, J., Hansen, I., & Norenzayan, A. (2009). Religion and support for suicide attacks. *Psychological Science, 20*(5), 612-618.

Gino, F., Norton, M. I., & Ariely, D. (2010). The counterfeit self: The deceptive costs of faking it. *Psychological Science, 21*(5), 712-720.

Gollwitzer, P. M., Sheeran, P., Michalski, V., & Seifert, A. E. (2009). When intentions go public: Does social reality widen the intention-behavior gap? *Psychological Science, 20*(2), 224-230.

Hong, S. W., & Shevell, S. K. (2009). Color-binding errors during rivalrous suppression of form. *Psychological Science, 20*(9), 1084-1091.

Jarvstad, A., Rushton, S. K., Warren, P. A., & Hahn, U. (2012). Knowing when to move on: Cognitive and perceptual decisions in time. *Psychological Science, 23*(6), 589-597.

Jiga-Boy, G. M., Clark, A. E., & Semin, G. R. (2010). So much to do and so little time: Effort and perceived temporal distance. *Psychological Science, 21*(12), 1811-1817.

Job, V., Dweck, C. S., & Walton, G. M. (2010). Ego depletion—Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological Science, 21*(11), 1686-1693.

Jonikaitis, D., & Deubel, H. (2011). Independent allocation of attention to eye and hand targets in coordinated eye-hand movements. *Psychological Science, 22*(3), 339-347.

Jostmann, N. (2013, March 22). Clipboard weight did not affect cogn. elaboration. Retrieved 06:06, January 03, 2014 from http://www.PsychFileDrawer.org/replication.php?attempt=MTU0

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*(5), 345-353.

Langer, E., Djikic, M., Pirson, M., Madenci, A., & Donohue, R. (2010). Believing is seeing: Using mindlessness (mindfully) to improve visual acuity. *Psychological Science, 21*(5), 661-666.

Lenth, R. V. (2009). Java applets for power and sample size [Computer software]. Retrieved June 22, 2013, from http://www.stat.uiowa.edu/~rlenth/Power.

Levelt, Noort & Drenth Committees. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel.* Downloaded from https://www.commissielevelt.nl

Linkenauger, S. A., Ramenzoni, V., & Proffitt, D. R. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science, 21*(12), 1854-1862.

Logan, G. D., & Crump, M. J. C. (2009). The left hand doesn't know what the right hand is doing: The disruptive effects of attention to the hands in skilled typewriting. *Psychological Science, 20*(10), 1296-1300.

Looser, C. E., & Wheatley, T. (2010). Illusory shrinkage and growth: Body-based rescaling affects the perception of size. *Psychological Science, 21*(9), 1318-1325.

Lupyan, G., Thompson-Schill, S. L., & Swingley, D. (2010). Conceptual penetration of visual processing. *Psychological Science, 21*(5), 682-691.

McGraw, A. P., Larsen, J. T., Kahneman, D., & Schkade, D. (2010). Comparing gains and losses. *Psychological Science, 21*(10), 1438-1445.

Noguchi, K., Handley, I. M., & Albarracín, D. (2011). Participating in politics resembles physical activity: General action patterns in international archives, United States archives, and experiments. *Psychological Science, 22*(2), 235-242.

Pratt, J., Radulescu, P. V., Guo, R. M., & Abrams, R. A. (2010). It's alive! Animate motion captures visual attention. *Psychological Science, 21*(11), 1724-1730.

Raffray, C. N., & Pickering, M. J. (2010). How do people construct logical form during language comprehension? *Psychological Science, 21*(8), 1090-1097.

R Development Core Team. (2011) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Rule, N. O., Rosen, K. S., Slepian, M. L., & Ambady, N. (2011). Mating interest improves women's accuracy in judging male sexual orientation. *Psychological Science, 22*(7), 881-886.

Sussman, A. B., & Shafir, E. (2012). On assets and debt in the psychology of perceived wealth. *Psychological Science, 23*(1), 101-108.

Swann, W. B., Jr., Gómez, Á., Dovidio, J. F., Hart, S., & Jetten, J. (2010). Dying and killing for one's group: Identity fusion moderates responses to intergroup versions of the trolley problem. *Psychological Science, 21*(8), 1176-1183.

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science, 22*(2), 253-260.

Vul, E., & Rich, A. N. (2010). Independent sampling of features enables conscious perception of bound objects. *Psychological Science, 21*(8), 1168-1175.

Waytz, A., & Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science, 23*(1), 77-85.

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science, 22*(11), 1434-1441.

Zwann, R. (2013, October 22). "Effects are public property and not personal belongings": A post-publication conversation. *A Candid Blog about Psychological Experimentation, Cognition, Language, and Other Stuff*. Retrieved January 03, 2014 from http://rolfzwaan.blogspot.nl/2013/10/effects-are-public-property-and-not.html.