Evaluating the Evidence for Enclothed Cognition: Z-Curve and Meta-Analyses

Abstract

Enclothed cognition refers to the systematic influence that clothes can have on the wearer's feelings, thoughts, and behaviors through their symbolic meaning. It has attracted considerable academic and non-academic interest, with the 2012 article that coined the phrase cited over 600 times and covered in over 160 news outlets. However, a recent high-powered replication failed to replicate one of the original effects. To determine whether the larger body of research on enclothed cognition possesses evidential value and replicable effects, we performed z-curve and meta-analyses using 105 effects from 40 studies across 24 articles (N = 3,789). Underscoring the marked improvement of psychological research practices in the mid-2010s, our results raise concerns about the replicability of early enclothed cognition studies but affirm the evidential value for effects published after 2015. These later studies support the core principle of enclothed cognition—what we wear influences how we think, feel, and act.

Keywords: enclothed cognition, embodied cognition, z-curve analysis, meta-analysis

Over a decade ago, Adam and Galinsky (2012) coined the term "enclothed cognition" to designate the systematic influence that clothes can have on the wearer's feelings, thoughts, and behaviors through their symbolic meaning. The goal was to offer a potentially unifying framework to integrate previous findings and explain the psychological impact of clothing on its wearers. Drawing from research on embodied cognition (e.g., Barsalou, 1999, 2008; Glenberg, 1997; Niedenthal, Barsalou, Winkielman, Krauth-Gruber, & Ric, 2005), the authors argued that it is the physical experience of wearing clothes combined with their symbolic associations that influences the wearer.

Results from three studies supported the enclothed cognition framework. In a first study, wearing a lab coat reduced errors during a Stroop test on incongruent trials compared to not wearing a lab coat. In two additional studies, wearing a lab coat described as a doctor's coat improved performance on a sustained attention task compared to wearing a lab coat described as a painter's coat and compared to merely seeing or identifying with a lab coat described as a doctor's coat.

Since its publication, the article on enclothed cognition has attracted considerable academic as well as non-academic interest: According to Google Scholar, it has been cited over 600 times; it has an Altmetric score of 1,890, which puts it in the top 0.02% of over 23 million tracked research outputs with 52 blog posts, 391 tweets, and 277 news stories from 163 outlets (https://www.altmetric.com/details/617583); and it has spawned over a dozen YouTube videos that have collectively accumulated more than a million views.

Recently, however, Burns, Fox, Greenstein, Olbright, and Montgomery (2019) failed to replicate the effect of wearing a lab coat on the Stroop test in a preregistered, high-powered direct replication attempt. This finding dovetails with the fact that replication failures have been

particularly prevalent in the domain of embodied cognition (e.g., Earp, Everett, Madva, & Hamlin, 2014; Garrison, Tang, & Schmeichel, 2016; Lynott et al., 2014), with failed replications including the links between body postures and power (Carney, Cuddy, & Yap, 2010), physical and interpersonal warmth (Williams & Bargh, 2008), and physical and moral cleansing (Zhong & Liljenquist, 2006). Taken together, failed replications have challenged the validity of embodied cognition in general and enclothed cognition in particular.

In their response to the failed replication by Burns et al. (2019), Adam and Galinsky (2019) acknowledged that the results cast doubt on the effect of wearing a lab coat on Stroop test performance. At the same time, they cited 20 studies from 10 articles that have conceptually replicated an enclothed cognition effect and concluded that "the sum total of the available data suggests that the core principle of enclothed cognition—what we wear can influence how we think, feel, and act—is generally valid" (p. 157). However, conceptual replications alone do not necessarily indicate support for a given finding if there is reason to believe that publication bias may be a concern (Earp & Trafimow, 2015; LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018); evidence of publication bias implies that for every conceptual replication that is published, one or more failed conceptual replications may have remained unpublished.

Furthermore, their review of the literature relied on a simple binary account of whether a study supports an enclothed cognition effect or not, which can be problematic for several reasons: First, because results typically get published only when they show an effect, the fact that a simple majority of published effects are statistically significant does not provide convincing evidence that an effect truly exists (Pashler & Harris, 2012; Simmons & Simonsohn, 2017). Moreover, a closer look at the articles cited in their review reveals that some of the referenced

studies are either correlational in nature or include potential confounds to the clothing manipulations that raise additional concerns about internal validity.

Given the impact of the original enclothed cognition paper on the one hand, and the replication failure of an enclothed cognition effect (as well as the multiple replication failures of broader embodied cognition effects) on the other hand, a more thorough analysis is needed. We therefore conducted a discerning review of the literature on enclothed cognition. Specifically, we performed a z-curve analysis to determine whether the extant literature provides reliable evidence for enclothed cognition effects and used various meta-analyses to estimate the underlying strength of these effects.

Z-Curve Analysis

Z-curve analysis was developed to assess the replicability of studies in a specific research area and to estimate whether published effects may be due to problematic research practices like *p*-hacking and publication bias (Bartoš & Schimmack, 2022). It relies on the premise that literatures plagued by questionable estimates are more likely to contain disproportionately large shares of *p*-values between .01 and .05 (i.e., values that have *just* reached statistical significance through misguided or inappropriate research practices) relative to non-significant values ($p > .05$) and "highly" significant values ($p < .001$). While theoretically similar to methods like *p*-curve analysis, z-curves are preferable when effects are heterogeneous because they provide more reliable estimates of replicability as well as additional information about the possible distribution of selection bias in a literature (Brunner & Schimmack, 2020). However, because z-

curves are a relatively recent innovation, we also report the better-known *p*-curve analysis in our Supplementary Material.[1]

Z-curves provide several key insights. First, they can quickly reveal the presence of publication bias, which can be evident if the plotted curve shows a steep drop of frequencies from "just significant" to "just non-significant" values (Schimmack & Brunner, 2017). Second, z-curves provide estimates for four separate and informative statistical parameters: the observed discovery rate, the expected discovery rate, the expected replication rate, and the maximum false discovery risk. Given the importance of these four parameters, we explain each in more detail below.

The **observed discovery rate** is the percent of effects reported in a literature that are statistically significant. It is important to note that this rate is not necessarily an accurate reflection of true or false findings. This rate can be inflated by selection for significance, which occurs when researchers selectively report significant findings or when journals selectively publish significant findings at disproportionately high rates. For example, the observed discovery rate in psychology journals is somewhere around 90% (Motyl et al., 2017; Schimmack, 2020), despite research demonstrating that the percentage of replicable effects contained in social science is likely closer to 37% (Open Science Collaboration, 2015).

The **expected discovery rate**, in contrast, is an estimate of the average power of all studies that were conducted in a literature. This statistic uses the truncated distribution of significant *z*-scores ($z > 1.96$) to estimate how many non-significant *z*-scores are likely to exist (published or not) and what portion of the overall literature they should represent based on

[1] As shown in detail in our Supplementary Material, we followed the official user guide by Simonsohn, Nelson, and Simmons (2015) and found that different *p*-curves (i.e., selecting the first, last, lowest, or highest *p*-value presented for each study) consistently provided evidential value for an enclothed cognition effect.

random sampling error. In short, it estimates an answer to the question: In how many studies would we expect to see both significant and non-significant effects? As an estimate of power, the expected discovery rate is compared against the observed discovery rate to determine whether there is clear statistical evidence for questionable research practices. Findings are considered suspect if an observed discovery rate is much higher than an expected discovery rate. When this difference is significant, it suggests some portion of effects in a literature can be explained by questionable research practices like selection bias, publication bias, or *p*-hacking (Bartoš & Schimmack, 2022).

The **expected replication rate** is an estimate of the mean power of the statistically significant effects in a literature. This rate reflects the probability of observing the same statistically significant effect, in the same direction, given the same parameters (e.g., identical study designs, sample sizes, etc.). This estimate is more informative when interpreted alongside the **maximum false discovery risk,** an estimate of the maximum possible percentage of false positives that would be consistent with the observed data. Importantly, a literature with a high expected replication rate and a low maximum false discovery risk indicates that significant results are more likely to be replicable, i.e., "true positives."

Study Exclusions

To predetermine the sample for our analysis, we set the following rules in advance: We committed to including all English language, peer-reviewed, experimental studies that examine enclothed cognition, i.e., the intrapersonal effects of clothes on the wearer's feelings, thoughts, and/or behaviors through the clothes' symbolic meaning. This means our analysis includes studies focused on hypotheses that invoke the symbolic meaning of clothes (e.g., "participants will be more helpful when wearing a nurse outfit"), but not on hypotheses that merely invoke the

functional effects of clothes (e.g., “wool sweaters will make you feel warmer than silk shirts”). We chose to include an array of different types of clothing rather than one specific type of clothing because it provides a conservative test of the null hypothesis (the symbolic meaning of clothing has no impact on a wearer).

We excluded any studies that were not written in the English language or peer-reviewed, e.g., studies reported in undergraduate theses, doctoral dissertations, book chapters, or review articles. For example, Brinol, Petty, and Belding (2017) provided a literature review and described a study on the relationship between wearing non-prescription reading glasses and information processing, but the study has not been peer-reviewed, so it was not included in our analysis (their description also did not include any statistical results).

Furthermore, because an explicit tenet of enclothed cognition is that it necessitates the physical experience of actually wearing the clothes, we excluded studies that only explored the effects of imagining what it would be like to wear the clothes (e.g., Kwon, 1994; Tiggemann & Andrew, 2012) or the effects of what virtual avatars wearing virtual clothing (e.g., Cutright, Srna, & Samper, 2019: Studies 2 and 3; Peña, Hancock, & Merola, 2009).

We excluded correlational studies (e.g., Burger & Bless, 2017; Ellis & Jenkins, 2015; Karl, Hall, & Peluchette, 2013; Peluchette & Karl, 2007; Slepian, Ferber, Gold, & Rutchick, 2015: Studies 1, 2, and 5; Solomon & Schopler, 1982) to ensure that any observed effects on the dependent variables would be caused by the clothes and not the other way around. For instance, Ellis and Jenkins (2015) presented three studies on the relationship between wearing a watch and conscientiousness and punctuality. Although it is plausible that wearing a watch increases conscientiousness and punctuality, it is also plausible that conscientious and punctual people are more likely to wear a watch.

Among experimental studies, we excluded the following studies in which the clothing manipulation was confounded with another manipulation: In Singer, Brush, and Lublin (1965), dressing up vs. dressing down was confounded with a deindividuation manipulation of having vs. not having a name tag. In Study 1a of Gino, Norton, & Ariely (2010), wearing counterfeit vs. authentic eyeglasses was confounded with a preference manipulation for counterfeit vs. authentic eyeglasses. In White & Carlson (2016) as well as White et al. (2017), dressing up as a character like Batman or Dora vs. not dressing up was confounded with a self-distancing manipulation.

Because enclothed cognition strictly relates to the intrapersonal effects of clothing, we excluded the following studies in which interpersonal influences could have explained the effects of clothing manipulations: Kraus and Mendes (2014) investigated the effects of wearing upper-class vs. lower-class clothing on testosterone levels and negotiation performance in an interactive negotiation, but it is unclear if the effects are truly intrapersonal as the clothing manipulation may have elicited different behaviors from negotiation partners. Likewise, Jones et al. (2019) measured the influence of wearing lab coats on students' levels of science interest, recognition as a science person, science self-efficacy, and STEM career goals, but any effects of wearing lab coats may have stemmed from being treated differently by their classroom science teacher.

We excluded interview and ethnographic studies that relied exclusively on qualitative data (e.g., Adomaitis & Johnson, 2005; Ogle, Tyner, & Schofield-Tomschin, 2013; Rafaeli, Dutton, Harquail, & Mackie-Lewis, 1997).

One study that met the criteria above but did not provide sufficient information for inclusion was by Vickroy, Shaw, and Fisher (1982). The authors examined the interactive effects of temperature, task complexity, and clothing on task performance and satisfaction. They

predicted a reversing interaction between clothing and temperature but did not report condition sizes, simple effect statistics, or standard errors.

Z-value Inclusions

In total, 40 studies from 24 articles met our criteria. In these studies, we were interested in the main effects of wearing a particular item of clothing as specified by the authors. Within this grouping, a few studies had 2 x 2 or more complicated designs. In rare cases, the nature of the dependent variable necessitated a second factor, which was typically a within-subjects factor, to measure an enclothed cognition effect. For example, to test whether wearing a lab coat facilitates selective attention, both Adam and Galinsky (2012: Study 1) and Burns et al. (2019) contrasted performance on the Stroop test in incongruent and non-incongruent trials, requiring a 2 x 2 interaction analysis to test for an enclothed cognition effect. In these two cases, we included the interaction effect.

In other cases, the authors included a second or third factor with predictions that it would not have any effect. For example, Cutright, Srna, and Samper (2019) expected that the effects of wearing something formal would hold across both high and low-prestige environments. Alternatively, some authors included a second or third factor and predicted an attenuated interaction effect. For instance, Dubois and Anik (2020) expected that the effects of wearing heels would be reduced in private rather than public settings. In these cases, the proper test of an enclothed cognition effect for the purposes of our analysis would be the main effect of wearing something formal or wearing heels, not the interaction with a moderating factor.

Finally, in some cases, hypotheses were specific for subgroups where one might expect the symbolic associations to differ. For example, Coyne et al. (2021) explicitly hypothesized that boys would be more prosocial when wearing “counter-gendered clothes” but made no prediction

for girls in their sample. Because it is possible that some cases like this, where main effects are only reported for subgroups, reflect post-hoc decisions to selectively highlight significant outcomes, we adopted two strategies to handle them. First, we included separate main effects as reported in the paper for the hypothesized subgroup. Second, in case the inclusion of any subgroup effects unduly influenced overall results, we ran models with and without their inclusion.

Data

In total, 105 estimates contained in 40 separate studies across 24 different articles were included in our z-curve -analysis (see Table 1). This data reflects 3,789 participants-worth of data (with a mean sample size of 94 participants per effect).

Coding

We used the "zcurve" package in R which requires *p*-values for every effect under consideration. These were obtained directly from papers or else estimated using effects and sample sizes reported in a given paper.[2] Note that five out of the 105 effects considered were originally coded as $p = 1$ because researchers did not provide any values (i.e., an effect was reported simply as "not significant"). These were included in our analysis below, but we observed no substantive differences when dropping them in supplementary analyses.

Table 1.

Effects in Z-Curve Analysis

(0) Original paper	(1) Study	(2) Methodology	(3) Dependent variable	(3) Key statistical result[3]	(5) Sample Size	(6) *P*-value	(7) Code

[2] In rare cases where authors only reported a *p*-value threshold without enough descriptive information to estimate accordingly, we conservatively default-coded to a more precise value (e.g., a non-significant finding would become $p = 1$, while $p < .001$ would become $p = .001$).

[3] Whenever authors of papers left out select statistics from their reporting that were required to calculate an effect size, we did our best to conservatively extrapolate based upon prior research. This was generally true of older papers. Extrapolation was

	# in Article						
Adam & Galinsky (2012)	1	lab coat vs. not wearing a lab coat (p. 920)	errors on incongruent trials of a Stroop task	Cohen's d = .63	n = 58	p = .02	attention
Adam & Galinsky (2012)	1	lab coat vs. not wearing a lab coat (p. 920)	response time on Stroop task	unreported (default to Cohen's d = 0)	n = 58	unreported (default to p = 1)	attention
Adam & Galinsky (2012)	2	lab coat vs. controls (i.e., wearing a painter's coat or looking at a lab coat) (p. 920)	the number of differences participants found on four comparative visual search tasks	Cohen's d = .78	n = 74	p = .02	attention
Adam & Galinsky (2012)	3	doctor's coat vs. controls (i.e., wearing a painter's coat or identifying with a doctor's coat) (p. 921)	the number of differences participants found on four comparative visual search tasks	Cohen's d = .87	n = 99	p < .001 (default to p = .001)	attention
Bailey, Horton, & Galinksy (2021)	1	home vs. work attire (p. 346)	self-reported authenticity	Cohen's d = .27	n = 177	p = .018	state
Bailey, Horton, & Galinksy (2021)	1	work vs. home attire (p. 346)	self-reported power	Cohen's d = .12	n = 177	p = .282	state
Bailey, Horton, & Galinksy (2021)	1	home vs. work attire (p. 346)	self-reported engagement	Cohen's d = .015	n = 177	p = .090	state
Bailey, Horton, &	2	home vs. work attire (p. 350)	self-reported authenticity	Cohen's d = .51	n = 116	p = .003	state

required for 14 estimates overall, with 13 of those estimates coming from research conducted before 2016. As an example of extrapolation, Hannover and Kuhnen (2002) did not provide any indication of standard deviation in their article, but we were able to estimate variance based on past research using the same task (i.e., having participants rate how much a list of adjectives described themselves; Cicero, Marin, Becker, & Kerns, 2016; Zickfeld, & Schubert, 2016). In this case, we divided the variance statistic obtained from past research by the sample size corresponding to the effect requiring extrapolation and then (erring on the conservative side) doubled this variance estimate *before* including it in our analysis. All meta-analysis models were run both with and without inclusion of these extrapolated estimates, and the results remained substantively unchanged.

Galinksy (2021)							
Bailey, Horton, & Galinksy (2021)	2	work vs. home attire (p. 350)	self-reported power	Cohen's *d* = -.23 (counter to prediction)	*n* = 116	*p* = .333	state
Bailey, Horton, & Galinksy (2021)	2	home vs. work attire (p. 350)	self-reported engagement	Cohen's *d* = .48	*n* = 116	*p* =.005	state
Burns, Fox, Greenstein, Olbright, & Montgomery (2019)	1	lab coat vs. no coat (p. 153)	errors on Stroop task	Cohen's *d* = .06	*n* = 200	*p* = .63	attention
Burns, Fox, Greenstein, Olbright, & Montgomery (2019)	1	lab coat vs. no coat (p. 153)	response time on congruent trials of a Stroop task	Cohen's *d* = .06	*n* = 200	*p* = .56	attention
Civile & Obhi (2017)	1	police uniform vs. mechanic uniform (p. 5)	reaction time seeing low-SES target	Cohen's *d* = .19	*n* = 28	*p* = .211	attention
Civile & Obhi (2017)	1	police uniform vs. mechanic uniform (p. 5)	reaction time seeing black targets	Cohen's *d* = .12	*n* = 28	*p* = .755	attention
Civile & Obhi (2017)	2	police uniform vs. mechanic uniform (p. 7)	reaction time seeing low-SES target	Cohen's *d* = -.41 (counter to prediction)	*n* = 28	*p* = .301	attention
Civile & Obhi (2017)	2	police uniform vs. mechanic uniform (p. 5)	reaction time seeing black targets	Cohen's *d* = -.03 (counter to prediction)	*n* = 28	*p* = .947	attention
Civile & Obhi (2017)	3	police uniform vs. exposure to police uniform (p. 9)	reaction time seeing low-SES target	Cohen's *d* = .26	*n* = 56	*p* = .328	attention
Civile & Obhi (2017)	3	police uniform vs. exposure to police uniform (p. 9)	reaction time seeing black targets	Cohen's *d* = .27	*n* = 56	*p* = .333	attention
Coyne, Rogers,	1	boys in gender-conforming superhero outfits	expressed preference for gender	Cohen's *d* = -.44	*n* = 65	*p* = .085	behavior

Shawcroft, & Hurst (2021)		vs. counter-gender princess outfits (p. 305)	conforming toys	(counter to prediction)			
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	boys in counter-gender princess outfits vs. gender-conforming superhero outfits (p. 305)	expressed preference for counter-gender toys	Cohen's *d* = .55	*n* = 65	*p* = .032	behavior
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	boys in counter-gender princess outfits vs. gender-conforming superhero outfits (p. 305)	prosocial behavior (number of knocked over pencils picked up)	Cohen's *d* = .64	*n* = 65	*p* = .013	behavior
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	boys in counter-gender princess outfits vs. gender-conforming superhero outfits (p. 305)	prosocial measure of reaction time before helping	Cohen's *d* = *.54*	*n* = 65	*p* = .035	attention
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	boys in gender-conforming outfits vs. (p. 305)	persistence on maze task	Cohen's *d* = -.13 (counter to prediction)	*n* = 65	*p* = .614	behavior
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	girls in gender-conforming princess outfits vs. counter-gender superhero outfits (p. 305)	expressed preference for gender conforming toys	Cohen's *d* = .27	*n* = 76	*p* = .236	behavior
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	girls in counter-gender superhero outfits vs. gender-conforming princess outfits (p. 305)	expressed preference for counter-gender toys	Cohen's *d* = -.34 (counter to prediction)	*n* = 76	*p* = .146	behavior
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	girls in counter-gender superhero outfits vs. gender-conforming princess outfits (p. 305)	prosocial behavior (number of knocked over pencils picked up)	Cohen's *d* = .13	*n* = 76	*p* = .581	behavior
Coyne, Rogers, Shawcroft, & Hurst (2021)	1	girls in counter-gender superhero outfits vs. gender-conforming princess outfits (p. 305)	prosocial measure of reaction time before helping	Cohen's *d* = .12	*n* = 76	*p* = .597	attention

Coyne, Rogers, Shawcroft, & Hurst (2021)	1	girls in counter-gender superhero outfits vs. gender-conforming princess outfits (p. 305)	persistence on maze task	Cohen's *d* = .08	*n* = 76	*p* = .743	behavior
Cutright, Srna, & Samper (2019)	1	formal vs. casual dress (p. 389)	confidence measured in self report	Cohen's *d* = .76	*n* = 294	*p* < .0001 (default to *p* = .0001)	state
Cutright, Srna, & Samper (2019)	1	formal vs. casual dress (p. 389)	confidence measured in actual money spent	Cohen's *d* = .25	*n* = 294	*p* = .038	behavior
Cutright, Srna, & Samper (2019)	1	formal vs. casual dress (p. 389)	confidence measured in number of items actually purchased	Cohen's *d* = .19	*n* = 294	*p* = .115	behavior
Dubois & Anik (2020)	3	heels vs. flats (p. 20)	self reported power	Cohen's *d* = .60	*n* =119	*p* = .002	state
Dubois & Anik (2020)	3	heels vs. flats (p. 20)	action orientation	Cohen's *d* = .45	*n* =119	*p* = .017	state
Dubois & Anik (2020)	3	heels vs. flats (p. 20)	abstract thinking	Cohen's *d* = .52	*n* =119	*p* < .01 (default to *p* = .01)	state
Dubois & Anik (2020)	3	heels vs. flats (p. 20)	self report of feeling attractive	Cohen's *d* = .90	*n* =119	*p* < .01 (default to *p* = .01)	state
Frank & Gilovich (1988)	4	black vs. white team uniforms (p.82)	increased preference for aggressive activities	Cohen's *d* = .76	*n* = 72	*p* = .0003	state
Frank & Gilovich (1988)	4	black vs. white team uniforms (p.82)	aggressive themes used in a Thematic Apperception Test	Cohen's *d* = .10	*n* = 72	*p* = .677	behavior
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	1	crew-neck sweater for women vs. one-piece swimsuit (p. 274)	number of cookies eaten	unreported (default to Cohen's *d* = 0)	*n* = 72	unreported (default to *p* = 1)	behavior

Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	swimsuit vs. sweater all genders (p. 277)	wrote about body shape and size	Cohen's d = .75	n = 82	p = .001	behavior
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	one-piece swimsuit vs. V-neck sweater for women (p. 277)	body shame	Cohen's d = .86	n = 42	p = .008	state
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	swim trunks vs. crew-neck sweater for men (p. 277)	body shame	Cohen's d = -.33	n = 40	p = .303	state
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	swimsuit vs. sweater all genders (p. 277)	feeling ashamed, humiliated, and disgraced	Cohen's d = .85	n = 82	p < .001 (default to p = .001)	state
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	swimsuit vs. sweater all genders (p. 277)	feeling repentant, guilty, and blameworthy	Cohen's d = .41	n = 82	p < .05 (default to p = .05)	state
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	swimsuit vs. sweater all genders (p. 277)	feeling silly, awkward, and foolish	Cohen's d = .50	n = 82	p < .05 (default to p = .05)	state
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	sweater vs. swimsuit all genders (p. 277)	amount of candy eaten	unreported (default to Cohen's d = 0)	n = 82	unreported (default to p = 1)	behavior
Fredrickson, Roberts, Noll, Quinn, & Twenge (1998)	2	sweater vs. swimsuit all genders (p. 277)	number of math problems solved	unreported (default to Cohen's d = 0)	n = 82	unreported (default to p = 1)	behavior
Gamble & Walker (2016)	1	bicycle helmet vs. baseball cap (p. 291)	risk-taking across trails of a simulated balloon inflation task	Cohen's d = .59	n = 80	p = .01	behavior
Gamble & Walker (2016)	1	bicycle helmet vs. baseball cap (p. 291	scale measure of sensation seeking	Cohen's d = .73	n = 80	p = .002	state

Gamble & Walker (2016)	1	bicycle helmet vs. baseball cap (p. 291)	state anxiety	Cohen's *d* = .09	*n* = 80	*p* = .66	state
Gino, Norton, Ariely (2010)	1b	counterfeit vs. authentic brand sunglasses (p. 716)	lied about performance on a matrix task	Cohen's *d* = .95	*n* = 91	*p* = .0001	behavior
Gino, Norton, Ariely (2010)	2	counterfeit vs. authentic brand sunglasses (p. 716)	reported acquaintances were more likely to behave dishonestly than	Cohen's *d* = .65	*n* = 79	*p* = .005	other
Gino, Norton, Ariely (2010)	2	counterfeit vs. authentic brand sunglasses (p. 716)	interpreted common excuses as less likely to be truthful	Cohen's *d* = .46	*n* = 79	*p* = .046	state
Gino, Norton, Ariely (2010)	2	counterfeit vs. authentic brand sunglasses (p. 716)	judged targets as more likely to behave dishonestly	Cohen's *d* = 1.72	*n* = 79	*p* < .0001 (default to *p* = .0001)	other
Gino, Norton, Ariely (2010)	3	counterfeit vs. authentic brand sunglasses (p. 718)	lied about performance on a matrix task	Cohen's *d* = .79	*n* = 100	*p* = .001	behavior
Gino, Norton, Ariely (2010)	3	counterfeit vs. authentic brand sunglasses (p. 718)	lower state authenticity	Cohen's *d* = .84	*n* = 100	*p* = .0001	state
Hannover & Kuhnen (2002)	1	formal vs. casual outfits (p. 2516)	endorse more formal trait adjectives (e.g., strategic and restrained) as self-descriptors	Cohen's *d* = .83	*n* = 60	*p* = .004	state
Hannover & Kuhnen (2002)	1	formal vs. casual outfits (p. 2516)	endorse less relaxed trait adjectives (e.g., emotional and easygoing) as self-descriptors	Cohen's *d* = .42	*n* = 60	*p* = .137	state

Hannover & Kuhnen (2002)	1	formal vs. casual outfits (p. 2516)	response latencies for formal adjectives	Cohen's *d* = .19	*n* = 60	*p* = .499	attention
Hannover & Kuhnen (2002)	1	formal vs. casual outfits (p. 2516)	response latencies for relaxed adjectives	Cohen's *d* = .13	*n* = 60	*p* = .653	attention
Hebl, King, & Lin (2004)	1	one piece swimsuit (for women) or speedo (for men) vs. sweater (p. 1326)	wrote about body shape and size	Cohen's *d* = .85	*n* = 400	*p* < .0001 (default to *p* = .0001)	behavior
Hebl, King, & Lin (2004)	1	one piece swimsuit (for women) or speedo (for men) vs. sweater (p. 1326)	body shame	Cohen's *d* = .23	*n* = 400	*p* = .032	state
Hebl, King, & Lin (2004)	1	one piece swimsuit (for women) or speedo (for men) vs. sweater (p. 1326)	self esteem	Cohen's *d* = .19	*n* = 400	*p* = .053	state
Hebl, King, & Lin (2004)	1	sweater vs. one piece swimsuit (for women) or speedo (for men; p. 1326)	number of math problems solved	Cohen's *d* = .29	*n* = 400	*p* = .005	behavior
Hebl, King, & Lin (2004)	1	sweater vs. one piece swimsuit (for women) or speedo (for men; p. 1326)	amount of candy eaten	Cohen's *d* = -.16 (counter to prediction)	*n* = 400	*p* = .21	behavior
Ishii, Numazaki, & Tado'oka (2019)	1	blue vs. pink for men with low self esteem (p. 136)	gender-related cognition measured on an Implicit Association Test	Cohen's *d* = 1.28	*n* = 4	*p* = .021	attention
Ishii, Numazaki, & Tado'oka (2019)	1	blue vs. pink for men with high self esteem (p. 136)	gender-related cognition measured on an Implicit Association Test	Cohen's *d* = -1.03 (counter to prediction)	*n* = 4	*p* = .05	attention
Ishii, Numazaki, & Tado'oka (2019)	1	blue vs. pink for men with low self esteem (p. 136)	endorsement of "masculine" traits as self-descriptors	Cohen's *d* = .67	*n* = 4	*p* = .21	state

Ishii, Numazaki, & Tado'oka (2019)	1	blue vs. pink for men with high self esteem (p. 136)	endorsement of "masculine" traits as self-descriptors	Cohen's *d* = -1.74 (counter to prediction)	*n* = 4	*p* = .004	state
Ishii, Numazaki, & Tado'oka (2019)	1	pink vs. blue (p. 136)	egalitarian attitudes about "traditional" sex roles	Cohen's *d* = .20	*n* = 20	*p* = .66	state
Ishii, Numazaki, & Tado'oka (2019)	1	blue vs. pink (p. 136)	"benevolent" " sexist attitudes (e.g., women should be protected)	Cohen's *d* = .22	*n* = 20	*p* = .63	state
Ishii, Numazaki, & Tado'oka (2019)	1	blue vs. pink (p. 136)	"hostile"" sexist attitudes (e.g., women exaggerate problems at work)	unreported (default to Cohen's *d* = 0)	*n* = 20	unreported (default to *p* = 1)	state
Johnson & Downing (1979)	1	women in a robe that looked like a "Ku Klux Klannish" vs. a nurse's gown (p. 1534)	shocks administered to target	Cohen's d = .57	*n* = 60	*p* = .032	behavior
Kouchaki, Gino, & Jami (2014)	1a	heavy vs. light backpack (p. 416)	state guilt	Cohen's *d* = *1.19*	*n* = 30	*p* = .13	state
Kouchaki, Gino, & Jami (2014)	1b	heavy vs. light backpack (p. 417)	chose data entry task (i.e., self-punishment) over a (fun) puzzle task	Cohen's *d* = *.83*	*n* = 90	*p* < .001 (default to *p* = .001)	state
Kouchaki, Gino, & Jami (2014)	1c	heavy vs. light backpack (p. 418)	state guilt	Cohen's *d* = *.81*	*n* = 54	*p* = .061	state
Kouchaki, Gino, & Jami (2014)	1c	heavy vs. light backpack (p. 418)	choice of less "guilt inducing" snack	Cohen's *d* = *.65*	*n* = 54	*p* = .039	behavior
Kouchaki, Gino, & Jami (2014)	2	heavy vs. light backpack (p. 418)	state guilt	Cohen's *d* = *.65*	*n* = 51	*p* = .025	state

Kouchaki, Gino, & Jami (2014)	3	heavy vs. light backpack (p. 419)	state guilt	Cohen's *d* = *.55*	*n* = 71	*p* = .024	state
Kouchaki, Gino, & Jami (2014)	3	heavy vs. light backpack (p. 419)	less over-reporting on task	Cohen's *d* = *.54*	*n* = 71	*p* = .026	behavior
Kouchaki, Gino, & Jami (2014)	4	heavy vs. light backpack (p. 420)	fluency score (number of words divided by writing time) while writing an essay about guilt	Cohen's *d* = .54	*n* = 62	*p* = .084	behavior
Lasaleta & Loveland (2019)	append ix B	counterfeit vs. authentic brand sunglasses (p. 716	lower state authenticity	Cohen's *d* = .59	*n* = 35	*p* = .018	state
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	1	wearing nurse scrubs vs. identifying with nurse scrubs or wearing cleaner scrubs (p. 224)	empathic concern	Cohen's *d* = .91	*n* = 150	*p* = .001	state
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	1	wearing nurse scrubs vs. identifying with nurse scrubs or wearing cleaner scrubs (p. 224)	state of distress	Cohen's *d* = -.38 (counter to prediction)	*n* = 150	*p* = .09	state
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	1	wearing nurse scrubs vs. identifying with nurse scrubs or wearing cleaner scrubs (p. 224)	helping behavior in a game	Cohen's *d* = 1.21	*n* = 150	*p* < .0001 (default to *p* = .0001)	behavior
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	1	wearing nurse scrubs vs. identifying with nurse scrubs or wearing cleaner scrubs (p. 224)	faster reaction time to help	Cohen's *d* = .35	*n* = 150	*p* = .001	attention
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	2	wearing nurse scrubs vs. identifying with nurse scrubs, wearing cleaner scrubs, or identifying with cleaner's scrubs (p. 224)	empathic concern	Cohen's *d* = 1.21	*n* = 100	*p* < .0001 (default to *p* = .0001)	state

Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	2	wearing nurse scrubs vs. identifying with nurse scrubs, wearing cleaner scrubs, or identifying with cleaner's scrubs (p. 224)	state of distress	Cohen's *d* = -.20 (counter to prediction)	*n* = 100	*p* = .32	state
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	2	wearing nurse scrubs vs. identifying with nurse scrubs, wearing cleaner scrubs, or identifying with cleaner's scrubs (p. 224)	helping behavior	Cohen's *d* = .53	*n* = 100	*p* = .010	behavior
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	2	wearing nurse scrubs vs. identifying with nurse scrubs, wearing cleaner scrubs, or identifying with cleaner's scrubs (p. 224)	less distracted by egotistic words on Stroop task	Cohen's *d* = .21	*n* = 100	*p* = .62	attention
Lopez-Perez, Ambrona, Wilson, and Khalil (2016)	2	wearing nurse scrubs vs. identifying with nurse scrubs, wearing cleaner scrubs, or identifying with cleaner's scrubs (p. 224)	more distracted by altruistic words on Stroop task	Cohen's *d* = 1.61	*n* = 100	*p* = .001	attention
Martins, Tiggemann, & Kirkbride (2007)	2	speedo vs. sweater for men (p. 640)	state self-objectification measured on word stem completion task	Cohen's *d* = .49	*n* = 125	*p* = .007	state
Martins, Tiggemann, & Kirkbride (2007)	2	speedo vs. sweater for men (p. 640)	wrote about body shape and size	Cohen's *d* = .05	*n* = 125	*p* = .577	behavior
Martins, Tiggemann, & Kirkbride (2007)	2	speedo vs. sweater for men (p. 640)	wrote about physical appearance	Cohen's *d* = .16	*n* = 125	*p* = .376	behavior
Martins, Tiggemann, & Kirkbride (2007)	2	speedo vs. sweater for men (p. 640)	body shame	Cohen's *d* = .33	*n* = 125	*p* = .064	state

Martins, Tiggemann, & Kirkbride (2007)	2	speedo vs. sweater for gay men (p. 640)	amount of snack food eaten	Cohen's *d* = .57	*n* = 57	*p* = .035	
Martins, Tiggemann, & Kirkbride (2007)	2	speedo vs. sweater for heterosexual men (p. 640)	amount of snack food eaten	Cohen's *d* = -.39 (counter to prediction)	*n* = 68	*p* = .115	behavior
Mendoza & Parks-Stamm (2020)	1	police uniform vs. casual clothes (p. 2358)	more likely to shoot unarmed targets on a shooter task	Cohen's *d* = .44	*n* = 178	*p* = .006	behavior
Slepian, Ferber, Gold, & Rutchick (2015)	3	formal vs. casual clothing (p. 663)	abstract thinking measured with category inclusiveness of weak exemplars (e.g., a camel being an appropriate example of a vehicle)	Cohen's *d* = .82	*n* = 34	*p* = .022	state
Slepian, Ferber, Gold, & Rutchick (2015)	4	formal vs. casual clothing (p. 664)	abstract thinking measured by the difference in response times between local processing (e.g., a large letter composed of small Ls) and global processing (e.g., a large L composed of small letters)	Cohen's *d* = .61	*n* = 54	*p* = .029	attention
Wang, Wang, Lei, & Chao (2021)	1	business suits vs. casual clothes (p. 790)	selected healthier snack options when given a choice	Cohen's *d* = .81	*n* = 79	*p* = .0005	behavior
Wang, Wang, Lei, & Chao (2021)	1	business suits vs. casual clothes (p. 790)	consumed less calories in potato chips	Cohen's *d* = 1.12	*n* = 79	*p* = .00001	behavior

Zhong, Bohns, & Gino (2010)	2	sunglasses vs. clear glasses (p. 313)	selfish plays in a dictator game	Cohen's *d* = .57	*n* = 50	*p* = .049	behavior
Zhong, Bohns, & Gino (2010)	2	sunglasses vs. clear glasses (p. 313)	state anonymity	Cohen's *d* = .63	*n* = 83	*p* = .005	state
Zhong, Bohns, & Gino (2010)	3	sunglasses vs. clear glasses (p. 313)	selfish plays in a dictator game	Cohen's *d* = .61	*n* = 83	*p* = .007	behavior
Van Stockum & DeCaro (2014)	1	lab coat vs. no coat	insight problems solved	Cohen's *d* = -.26	*n* = 96	*p* = .555	behavior

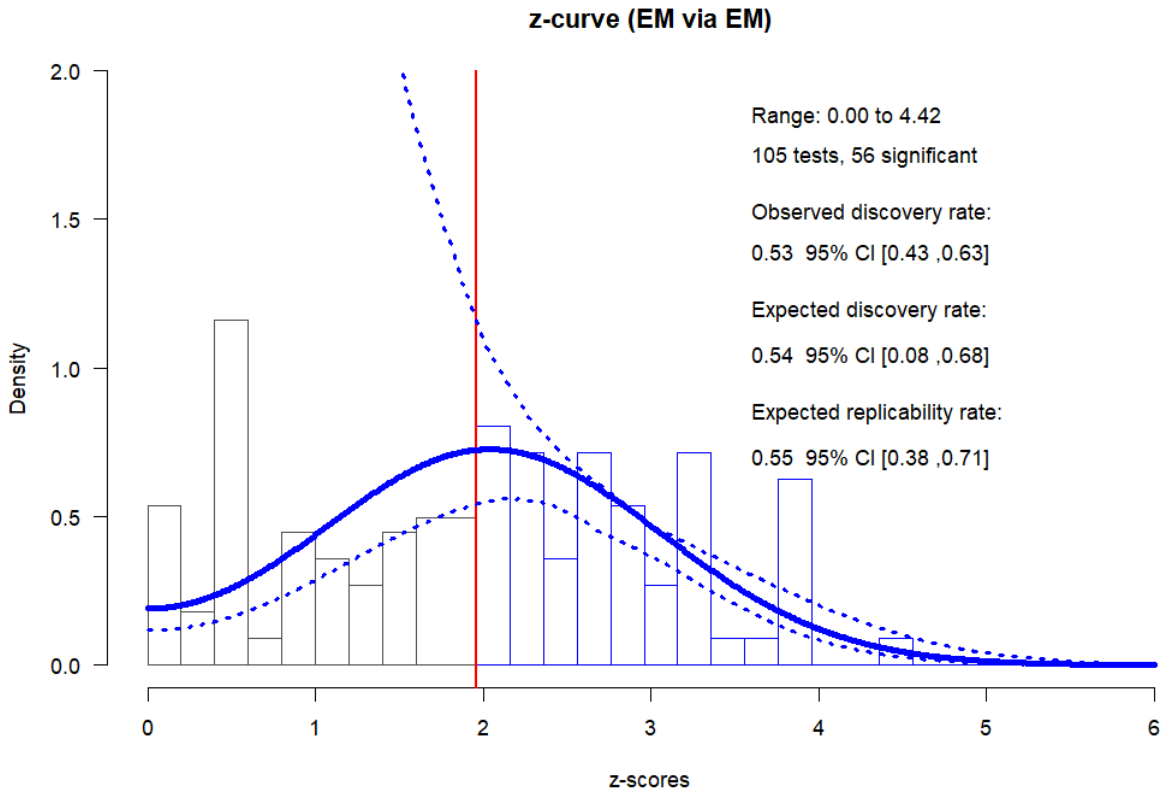
Note. *P*-values above relate to main effect estimates from the given statistical test reported in each article. If a *p*-value was not reported or only reported for an interaction model, we used *n* and *r* (often converted from another effect size) to calculate a *t*-value before looking up a corresponding two-tailed *p*-value. *P*-values are only reported to the second digit (rather than the third digit) if they appeared that way in their respective articles.

Results

All analyses were run using the "zcurve" package in R version 4.2.2. (Bartoš & Schimmack, 2020). We first analyzed all effects in our data (see Figure 1). A visual examination of the data provides anecdotal evidence for a small publication bias (seen by comparing bars directly to the left and right of the red line in the Figure below), though the distribution of non-significant findings in our data overall suggests this bias is not significant. The observed discovery rate (53%) was roughly equivalent to the expected discovery rate (54%) and fell inside the expected discovery rate's confidence interval (8% to 68%), which means this analysis did not provide clear evidence for questionable research practices. The expected replication rate was 55% with a confidence interval from 38% to 71%; and the false discovery risk was 5% with a confidence interval of 2% to 73%. These estimates suggest that a majority of significant results are likely to replicate under the same conditions. However, wide confidence intervals mean we cannot rule out the possibility of this pattern being explained by false positives.

Figure 1.

Z-Curve for All Enclothed Cognition Effects



Note. The histogram shows the distribution of observed z-statistics. The vertical red line shows the statistical significance criterion. The full blue line displays the density of the estimated model with the dotted lines reflecting confidence intervals.

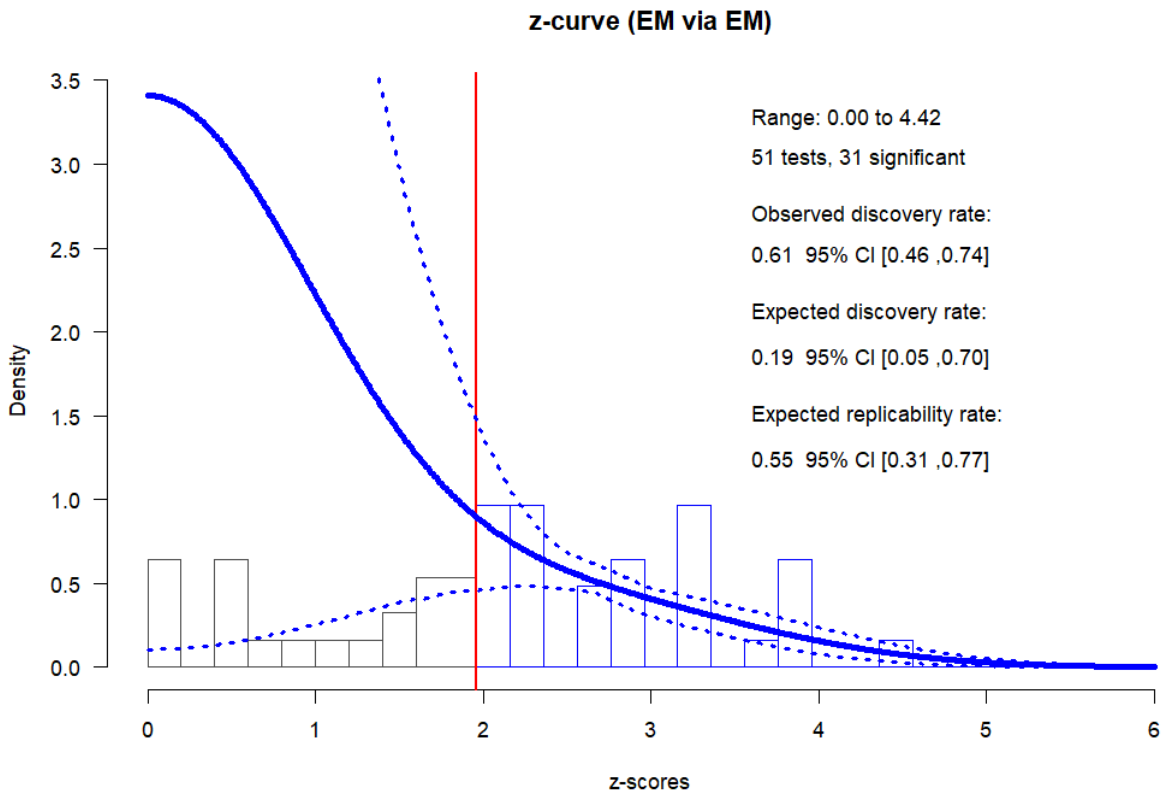
Because the past decade has brought about important changes to research practices designed to increase statistical and methodological rigor, we partitioned the data to compare older studies to more recent studies. Specifically, we compared studies performed before 2016 to ones performed 2016 and after. This threshold was selected because it marked a distinct increase in attention paid to both statistical power and replication; the Open Science Collaboration drew considerable attention to the replication crisis in 2015, psychology was facing heavy critiques of research practices featured in many well-known journals (Bohannon, 2015; Francis, 2014; Lindsay, 2015; Maxwell, Lau, & Howard, 2015), and data reporting practices across the social

sciences were changing (Brodeur, Lé, Sangnier, & Zylberberg, 2016; Christensen, & Miguel, 2018; Van't Veer & Giner-Sorolla, 2016). Moreover, methods like *p* and z-curves were introduced around that time, specifically to diagnose and respond to replication concerns (Simonsohn, Nelson, & Simmons, 2014; Schimmack, 2015). This partition produced an almost even split our data, with 51 effects reported before 2016 and 54 effects reported in 2016 and after.

Studies before 2016. Figure 2 plots the z-curve for effects *before* 2016. The observed discovery rate in these studies (61%) was substantially higher than the expected discovery rate (19%), though it still fell within the expected discovery rate's confidence interval (5% to 70%). The expected replication rate was again 55% with a confidence interval from 31% to 77%; and the false discovery risk was 23% with a confidence interval from 2% to 100%. Given the gap between the observed and expected discovery rates and the fact that the upper confidence interval for the false discovery risk is 100%, this analysis does not rule out the possibility that questionable research practices played a role in a portion of effects observed prior to 2016 or that a majority of significant findings are explained by false positives (Schimmack & Bartos, 2023).

Figure 2.

Z-Curve for Enclothed Cognition Effects Before 2016

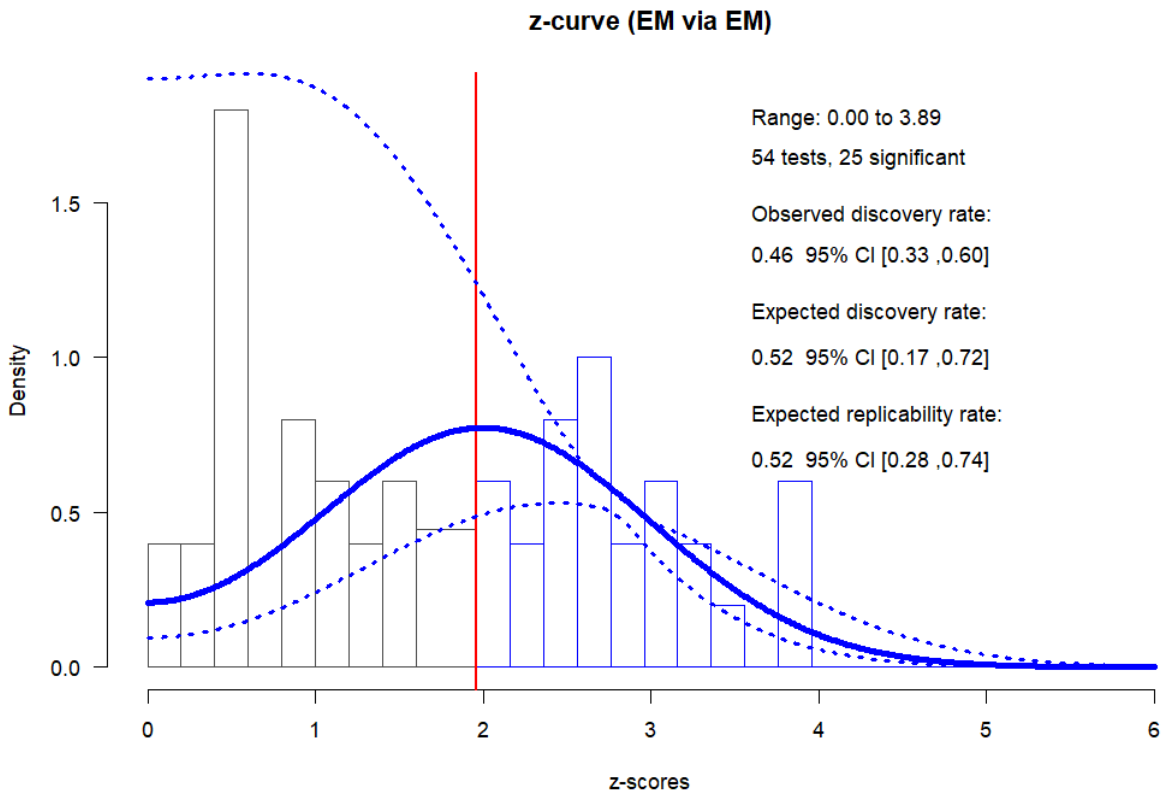


Note. The histogram shows the distribution of observed z-statistics. The vertical red line shows the statistical significance criterion. The full blue line displays the density of the estimated model with the dotted lines reflecting confidence intervals.

Studies in 2016 and after. Figure 3 plots a z-curve analysis for effects in *2016 and after*. The observed discovery rate in these studies (46%) was *lower* than the expected discovery rate (52%) and fell near the center of the expected discovery rate's confidence interval (17% to 72%). These results suggest that questionable research practices were unlikely to play a role in effects observed 2016 and after. What is more, the expected replication rate (52%) remained largely unchanged with a confidence interval from 47% to 72%; importantly, the false discovery risk (5%) was markedly smaller with a much narrower confidence interval from 2% to 6%. This pattern suggests that a majority of significant results from 2016 and after are likely to replicate under the same conditions and are unlikely to be explained by false positives.

Figure 3.

Z-Curve for Enclothed Cognition Effects in 2016 and After

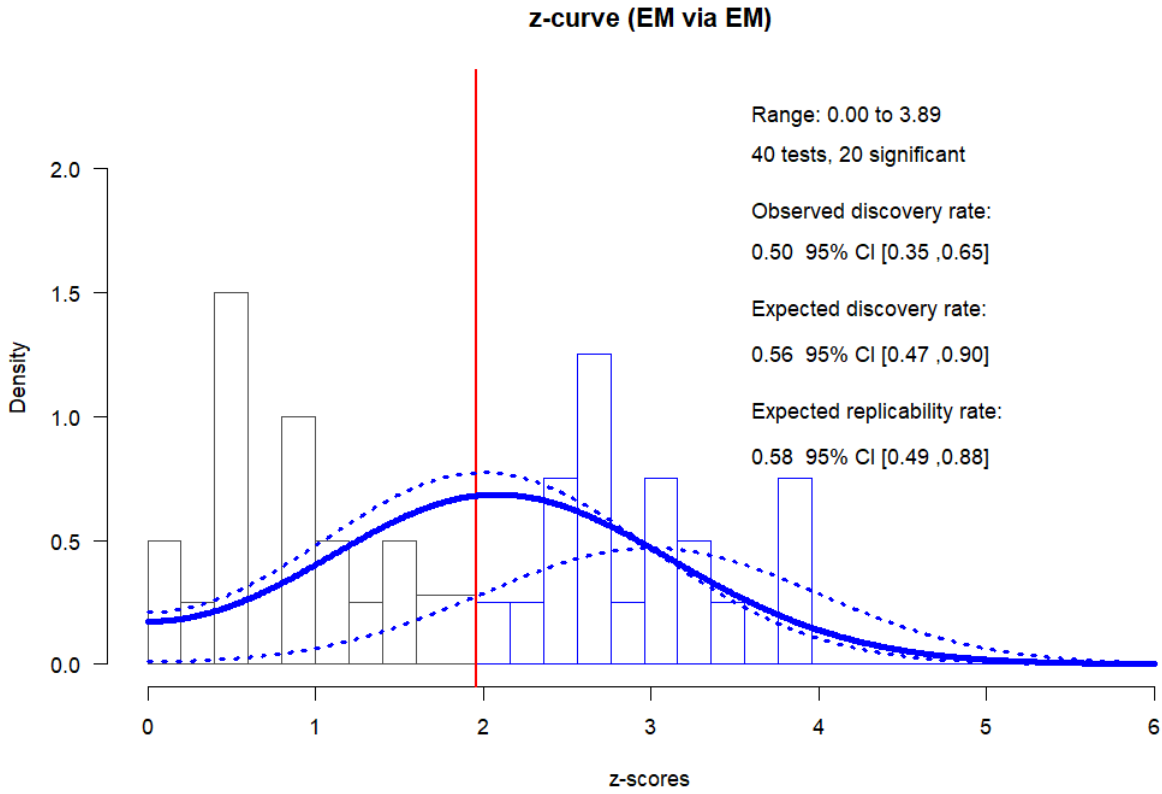


Note. The histogram shows the distribution of observed z-statistics. The vertical red line shows the statistical significance criterion. The full blue line displays the density of the estimated model with the dotted lines reflecting confidence intervals.

In some cases, effect sizes were only reported for subgroups without providing enough information to compose a precise estimate for an overall effect. For instance, Ishii, Numazaki, and Tado'oka (2019) tested the effect of blue vs. pink clothing on a sample of men, but only reported effects for men whose self-esteem was one standard deviation above and below the mean. Because reporting subgroup effects like this may reflect selective reporting practices, Figure 4 plots a z-curve analysis of studies conducted 2016 and after without the inclusion of subgroup estimates.

Figure 4.

Z-Curve for Enclothed Cognition Effects 2016 and after (Subgroup Effects Dropped)



Note. The histogram shows the distribution of observed z-statistics. The vertical red line shows the statistical significance criterion. The full blue line displays the density of the estimated model with the dotted lines reflecting confidence intervals.

Again, the observed discovery rate (50%) was *lower* than the expected discovery rate (56%). Notably, the observed discovery rate also fell at the lower end of the confidence interval for the expected discovery rate, which was both narrower and more positive (47% to 88%) than in our previous z-curves. This pattern raises the possibility that some portion of non-significant findings contained in this data may be attributable to false *negatives* observed in underpowered studies. Finally, the expected replication rate (58%) was slightly higher than in our previous analyses with a confidence interval from 47% to 88%; while the false discovery risk (4%) was smaller with a confidence interval from 1% to 6%. Again, this pattern suggests that a majority of the significant effects reported in these studies are likely to replicate under the same conditions and are unlikely to be explained by false positives.

Discussion

Our z-curve analyses yield several important insights. First, the overall body of work on enclothed cognition is likely to hold evidential value. Second, questionable research practices are likely to have played a role in studies conducted before recent methodological advances. However, these questionable practices are unlikely to have played a role in more recent studies. Furthermore, the observed discovery rate across all z-curves of 46% to 61% was lower than estimates for the observed discovery rate for research in the field of psychology of 90% (e.g., Motyl et al., 2017; Schimmack, 2020). In short, these results do not provide evidence that publication bias is a concern for research on enclothed cognition conducted in more recent years. Moreover, the expected replication rate for studies published before 2016 was higher than estimated rates for other social science studies before 2016 (e.g., The Open Science Collaboration estimated a replication rate of 37%; 2015), and the expected replication rate for studies published in 2016 and after was comparable to rates for top science journals today (e.g., Camerer et al., 2018 estimated a replication rate of 62% for *Nature* and *Science*). In sum, our analyses suggest the majority of enclothed cognition studies, especially those conducted in 2016 and after, are likely to replicate with sufficient power.

Meta-Analysis

One question that z-curve analyses cannot answer is how impactful enclothed cognition effects are. While z-curves can be used to estimate distributions of significant effects and detect the possibility of publication bias or *p*-hacking, they do not provide information about the *size* of an underlying effect, nor do they account for unique variation attributable to multiple estimates reported within a single study (Bartoš & Schimmack, 2022; Simonsohn et al., 2014; 2015). To address these limitations, we conducted a random-effects meta-analysis for all effects contained

in our data (n = 105) as well as for those studies which our z-curve analysis suggested are more likely to contain evidential, i.e., studies performed in 2016 and after not including subgroups (n = 40). In addition, because it is useful to know how enclothed cognition effects may vary depending upon the outcome under consideration, we include supplemental meta-analyses examining different *kinds* of dependent variables, i.e., behaviors, psychological states, and attention. We then use these analyses to make recommendations about adequate statistical power for future research.

Method

Data

We first conducted a meta-analysis on all the data used in our z-curve analyses. The data contained 105 estimates from 40 separate studies in 24 different articles, reflecting 3,789 participants-worth of data. We then conducted a second meta-analysis, only using estimates from our post-2015 z-curve. The data contained 40 estimates from 15 separate studies in 11 different articles, reflecting 1,693 participants-worth of data.

Coding

In addition to the effects and sample sizes from every study, we categorized different types of dependent variables into psychological state outcomes (e.g., authenticity, engagement, and anxiety), attentional outcomes (e.g., reaction-times, errors on a Stroop task, and cognitive focus or fixation), and behavioral outcomes (e.g., eating food, cheating on a test, and taking risks). In the full data set, 35.2% reflected psychological state outcomes, 25.7% reflected attentional outcomes, 34.3% reflected behavioral outcomes, and 4.8% were categorized as "other." In the post-2015 data set, 27.5% of the effects reflected psychological state outcomes, 20% reflected attentional outcomes, and 52.5% reflected behavioral outcomes.

Analysis

We used a random-effects approach to meta-analysis. This approach is conservative and used when the analytical aim is to generalize beyond available studies without assuming a single "fixed" effect size (Goh, Hall, & Rosenthal, 2016, p. 539). To compare the different types of estimates used in different studies, effect sizes were all converted to the same metric (Fisher's *z* transformation of the Pearson correlation coefficient) for our analysis, but outcomes are reported as Cohen's *d* for the reader's convenience. A complete list of these transformations with references for the formulas applied to different estimates is reported in Table 2 in the Supplemental Material.

First, we conducted a simple random effects meta-analysis using the "meta" package in R (Balduzzi, Rücker, & Schwarzer, 2007). Then, because many studies have more than one dependent variable with different corresponding effect sizes, we conducted a secondary robust variance estimation or "RVE" meta-analysis (Hedges, Tipton, & Johnson, 2010) using the "robumeta" package (Fisher, Tipton, & Hou, 2017). In recent years, RVE has become a popular meta-analytic technique employed as an alternative to simpler within-study averaging of effect sizes (Agadullina & Lovakov, 2018; Friese, Frankenbach, Job, & Loschelder, 2017; Kurdi et al., 2019). This is because RVE meta-analysis can better account for statistical dependency across effect sizes and often results in less information loss (Friese et al., 2017). Specifically, this method provides "a way to include all dependent effect sizes in a single meta-regression model, even when the exact form of the dependence is unknown" (Pustejovsky & Tipton, 2022, p. 1).[4]

Finally, to account for any potential publication bias, we further supplemented our investigation with Robust Bayesian Meta-Analysis with Publication Selection Model-Averaging

[4] All data and code for our meta-analyses can be found here: https://osf.io/2ytvn/?view_only=312fac17a53548cf80314a22a8c84fb2

(RoBMA-PSMA)—a particularly conservative meta-analytic technique that corrects for potential publication bias with estimates that are shrunken toward zero (Bartoš, Maier, Wagenmakers, Doucouliagos & Stanley, 2023, p. 109). This method incorporates a variety of selection models, works well when estimates show high heterogeneity, and includes precision-effect tests as well as precision-effect estimates with standard errors (PET-PEESE) that adjust for small-study effects by modeling the relationship between the effect sizes and standard errors (Carter et al., 2019; Bartoš et al., 2023). This form of meta-analysis provides Bayes Factor (BF_{10}) statistics that estimate the likelihood of heterogeneity and publication bias. This analysis was run using the "rjags" and "RoBMA" packages in R (Plummer, 2022; Bartoš, Maier, 2020).

Results

The simple random-effects mean effect size aggregated by study for all studies was d = .41, 95% CI = [.32 .49], $t(104) = 10.47$, $p < .0001$, $\tau = .03$, and for studies in 2016 and after was d = .41, 95% CI = [.28, .54], $t(39) = 6.42$, $p < .0001$, $\tau = .03$.[5] The RVE random-effects mean effect size with estimates aggregated by study for all studies was d = .47, 95% CI = [.39, .56], $t(34.5) = 9.97$, $p < .0001$, $\tau = .00$, and for studies in 2016 and after was d =.39, 95% CI = [.26, .52], $t(12.3) = 6.42$, $p < .0001$, $\tau = .00$.[6] We provide a complete list of individual effects and estimate weights for the RVE meta-analysis in Figures 3 and 4 in the Supplemental Material.

To assess enclothed cognition estimates on a more granular level, we conducted separate meta-analyses for different types of outcomes (see Table 2). Notably, the suggested confidence intervals appear relatively stable across outcomes, with the exception of attentional outcomes, which have a confidence interval that includes zero in the post-2015 data set.

[5] If we include subgroup main effects in this analysis, the results are d = .17, 95% CI = [.12, .23], $t(53) = 6.25$, $p <$ .0001.

[6] If we include subgroup main effects in this analysis, the results are d = .39, 95% CI = [.26, .52], $t(12.8) = 6.34$, $p <$ .0001.

Normality and representativeness were assessed by quantifying the proportion of effects above and below the meta-analytic means (Mathur & VanderWeele, 2019). This relatively simple test can help surface non-normality because when two-thirds of effects are *below* the meta-analytic mean it suggests an estimate has likely been inflated by disproportionately positive outliers or the selective suppression of unfavorable results (Bakdasha, & Marusich, 2022; Formann, 2008). This metric suggests that evidence of non-normality or bias is not an issue for psychological states or behavioral outcomes, but it does raise concerns about biased estimation for attentional outcomes in the post-2015 data set (see Table 2).

Table 2.

Effect Sizes Provided by RVE-Analysis and Sub-Group RVE-Analyses (All Studies)

Type of Dependent Variable	Estimated Effect Size (Cohen's *d*) with 95% Confidence Intervals	T-value and Degrees of Freedom	P-value	Tau Squared	Proportion of Effects Below the RVE Meta-Analytic Mean
Overall	.47 [.39, .55]	$t(34.5) = 9.97$	$p < .0001$	.00	*49.52%*
Behavior	.47 [.31, .65]	$t(17.2) = 5.58$	$p < .0001$	.00	*50.00%*
State	.52 [.37 .68]	$t(15.4) = 7.79$	$p < .0001$	.00	*45.95%*
Attention	.43 [.18, .68]	$t(12.2) = 3.93$	$p = .002$	.00	*62.96%*
Other	.35 [.35, .66]	$t(1.99) = 4.8$	$p = .04$	.00	*60.00%*
Overall	.39 [.26, .52]	$t(12.3) = 6.44$	$p < .0001$	.00	*52.50%*
Behavior	.61 [.26, 1.08]	$t(4.66) = 3.79$	$p = .015$	.00	*50.00%*
State	.43 [.26 .61]	$t(7.17) = 4.78$	$p = .002$	.00	*47.62%*

Attention	.32 [0, .68]	$t(4.3) = 2.02$	$p = .11$	.00	*81.82%*

Note. Meta-analytic results can become less reliable when the degrees of freedom approach four (as is true of the "other" category above, as well as post-2015 behavioral and attentional outcomes), though some scholars suggest the use of lower p-values (e.g., $p < .01$) can mitigate this concern (Tanner-Smith, Tipton, & Polanin, 2016).

Supplementary Robust Bayesian meta-analysis using all effects showed a small mean model-averaged estimate of Cohen's $d = .046$, 95% CI [0.00, 0.33] (with the lower bound above zero), evidence of heterogeneity ($BF_{rf} = 9.997$) with a mean model-averaged estimate $\tau = .390$, 95% CI [.30, .48], and strong evidence of publication bias ($BF_{pb} = 127.19$). In contrast, post-2015 effects showed a higher mean model-averaged estimate of Cohen's $d = .259$, 95% CI [0.00, 0.45] (with the lower bound above zero), evidence of heterogeneity ($BF_{rf} = 5.27$) with mean model-averaged estimate $\tau = .360$, 95% CI [0.26, 0.50], and no evidence of publication bias ($BF_{pb} = .725$).[7] It should be noted that because RoBMA-PSMA estimates incorporate multiple models that are designed to correct for publication bias (i.e., some models assume publication bias and some do not), its estimates are shrunken toward zero and come with wider credible intervals by design. This is meant "to reflect the additional uncertainty about the publication bias process" (Bartos et al., 2023, p. 6). Taken together, these results echo the results from our z-curves and standard meta-analyses, with a more pronounced contrast between studies published before 2016 and studies published in 2016 and after. While the earlier studies appear to suffer from publication bias and do not offer support for an enclothed cognition effect, the later studies do not appear to suffer from publication bias and support a moderate enclothed cognition effect.

[7] In contrast, pre-2016 effects showed a mean model-averaged estimate Cohen's $d = .033$, 95% CI [-0.08, 0.32], evidence of heterogeneity, $BF_{rf} = 1.11$, with mean model-averaged estimate $\tau = 0.384$, 95% CI [0.27, 0.52], and strong evidence of publication bias, $BF_{pb} = 471.51$.

Finally, as an additional robustness check we reran meta-analyses (simple, RVE, and RoBMA-PSMA) on effects observed in 2016 publications and later for only those studies which used cover stories (n = 32). We conducted these analyses to account for the possibility of demand characteristics (i.e., participants correctly guessing a study's hypothesis and subsequently performing in ways that fulfill hypothesized effects). Demand characteristics do not appear to be a concern as 60% of the effects from studies with cover stories were significant, as compared to 27% of the effects from studies without cover stories ($\chi^2 = 4.19,\ p = .04$). In addition, the direction and significance of overall estimated effects remained largely unchanged when only considering effects from studies that used a cover story (simple Cohen's d = .48, 95% CI [0.32, 0.65]; RVE Cohen's d = .47, 95% CI [0.28, 0.68]; and RoBMA-PSMA Cohen's d = .34, 95% CI [0.00, 0.60], with evidence of heterogeneity($BF_{rf} > 100$) with mean model-averaged estimate τ = 0.401, 95% CI [0.26, 0.63]), and weak evidence of publication bias (BF_{pb} = 1.06).

Discussion

The goals of the current research were to (a) assess whether the literature on enclothed cognition possesses evidential value and (b) estimate an effect size for any studies that possess evidential value. In layman's terms, do symbolic associations with clothes influence the wearer's feelings, thoughts, and behaviors? And, if so, to what extent? To answer these questions, we conducted both z-curve and meta-analyses.

Our z-curve analyses suggest that enclothed cognition research contains evidential value. Moreover, although separate z-curves split by publication year cannot conclusively rule out the possibility of questionable research practices in studies conducted before 2016, they demonstrate that there is little evidence for these practices in more recent work. We attribute this increase in

reliability to increased attention and effort dedicated to assessing and ensuring methodological rigor since the mid-2010s.

We subsequently conducted simple and RVE meta-analyses on all studies as well as studies published in 2016 and later. These analyses point to reliable effect sizes in the small-to-medium range. Effects were consistent and robust across different types of dependent variables (e.g., psychological states and behaviors), often ranging from $d = .32$ to $d = .61$.[8] This is consistent with average effect sizes of around $d = .43$ ($r = .21$) estimated for social psychology studies more generally (Richard, Bond, and Stokes-Zoota, 2003). What is more, results remained relatively consistent, even when only analyzing studies that employed cover stories, suggesting demand characteristics are unlikely to explain the estimated effect sizes. Finally, the results from our additional RoBMA-PSMA meta-analysis largely align with our z-curve analysis, albeit with a more critical view of older studies. Although there was evidence of publication bias for studies before 2016, there was no indication of publication bias for studies in 2016 and after. These results suggest a conservative estimate of $d = .26$ (for effects in 2016 and after) is more likely to be reliable.

As Maxwell and colleagues note in their review of the replication crisis, "it remains to be seen how many of the recently claimed failures to replicate will be supported or instead may turn out to be artifacts of inadequate sample sizes and single study replications" (2015, p. 1). To address this concern directly, we provide one final recommendation based on supplementary analyses. Although statistically significant effects in our data were correlated with larger sample

[8] One caveat is that in the 2016 and later data set, attentional outcomes showed the weakest effects in our data with the confidence interval approaching zero and evidence of non-normality. But it is also worth noting that the average sample size for effects attached to attention outcomes ($n = 79$) was smaller than the average sample sizes attached to either state ($n = 112$) or behavioral outcomes ($n = 114$).

sizes,[9] most studies in the enclothed literature to date remain underpowered. The average overall sample size when calculating between-subjects two-group effects was roughly 103 participants, which (using the "pwr" package in R) corresponds to 50% power using the RVE meta-analytic mean suggested by studies published in 2016 and later (d = .39) and only 25% power using our estimated RoBMA-PSMA meta-analytic mean suggested by studies published in 2016 and later (d = .26). Power analysis suggests that studies seeking to replicate these effects would require total samples between 234 to 770 participants to obtain 95% power with a two-group design. Beyond using these thresholds for future power analysis, we recommend a straightforward minimum of 150 participants per condition as a useful heuristic for researchers seeking to test, extend, or replicate enclothed cognition effects. We make this recommendation for testing *main* effects, as sample sizes would be appreciably higher when seeking to test more complex models and interactions.

Finally, many of the studies in our data relied exclusively on samples of Western participants observed in laboratory studies. Our analysis cannot speak to how different cultural contexts might affect enclothed cognition effects more broadly, though it is logical to assume that the symbolic value of clothing varies depending upon culture and context. For example, Bailey et al. (2022) found that wearing "home clothes" when at home led remote workers to feel more authentic and engaged during their workday. Similarly, garments imbued with national or religious meaning (like wearing a headscarf) may vary depending on the cultural settings they are worn in (in a temple or on the street). An examination of cultural factors is surprisingly absent from the current literature but provides promising new avenues for future research.

[9] A t-test comparing sample sizes attached to effects that were and were not significant in our meta-analysis showed that significant effects were, on average, more prevalent in studies with more participants (n = 127) versus fewer participants (n = 79), $t(38.74) = 2.53$, $p = .02$.

Conclusion

Assessing the evidential value of the current literature is critical for interpreting past research and to guide future research, especially in light of the replication crisis surrounding social psychology more broadly and embodied cognition more specifically (Maxwell, Lau, & Howard, 2015; Wilson & Lipsey, 2001). To that end, we conducted both z-curve and meta-analyses to test the reliability and effect size of enclothed cognition effects. Although our analyses support concerns about publication bias in early research on enclothed cognition, we found reliable evidence for enclothed cognition in more recent, methodologically rigorous studies—even after accounting for alternative explanations like demand characteristics. Given the impact that enclothed cognition has made in academic circles, news outlets, and social media platforms on the one hand, and the recent replication failures of enclothed cognition and embodied cognition effects on the other hand, these results are encouraging. In their original article on enclothed cognition, Adam and Galinsky (2012) opened with the following quote from Nobel Prize winning author Isaac Bashevis Singer: “What a strange power there is in clothing.” Our analyses suggest that Singer’s sentiment is scientifically warranted.

References

Adam, H., & Galinsky, A. D. (2012). Enclothed cognition. *Journal of Experimental Social Psychology*, *48*, 918-925.

Adam, H., & Galinsky, A. D. (2019). Reflections on enclothed cognition: Commentary on Burns et al. *Journal of Experimental Social Psychology*, *83*, 157-159.

Adomaitis, A. D., & Johnson, K. K. (2005). Casual versus formal uniforms: Flight attendants' self-perceptions and perceived appraisals by others. *Clothing and Textiles Research Journal*, *23*, 88-101.

Agadullina, E. R., & Lovakov, A. V. (2018). Are people more prejudiced towards groups that are perceived as coherent? A meta-analysis of the relationship between out-group entitativity and prejudice. *British Journal of Social Psychology*, *57*(4), 703-731.

Bailey, E. R., Horton, C. B., & Galinsky, A. D. (in press). Enclothed harmony or enclothed dissonance? The effect of attire on the authenticity, power, and engagement of remote workers. *Academy of Management Discoveries*.

Bakdasha, J. Z., & Marusich, L. R. (2022). Left-truncated effects and overestimated meta-analytic means. *Proceedings of the National Academy of Sciences*, *119*(31), e2203616119.

Balduzzi, S., Rücker, G., Schwarzer, G. (2019). "metafor. How to Perform a Meta-analysis with R: a Practical Tutorial, Evidence-Based Mental Health". 22: 153-160.

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and Brain Sciences, 22*, 577–609.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.

Bartoš F., Maier, M. (2020). "RoBMA: An R Package for Robust Bayesian Meta-Analyses." R

package version 2.3.2, <https://CRAN.R-project.org/package=RoBMA>.

Bartoš, F., Maier, M., Wagenmakers, E. J., Doucouliagos, H., & Stanley, T. D. (2023). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, 14(1), 99-116.

Bartoš, F., & Schimmack, U. (2020). "zcurve: An R package for Fitting Z-curves." R package version 2.3.0, <https://CRAN.R-project.org/package= zcurve>

Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology*, *6*.

Bohannon, J. (2015). Many psychology papers fail replication test. *Science*, 349(6251), 910–911. doi:10.1126/science.349.6251.910

Brinol, P., Petty, R. E., & Belding, J. (2017). Objectification of people and thoughts: An attitude change perspective. *British Journal of Social Psychology, 56*, 233-249.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1-32.

Burger, A. M., & Bless, H. (2017). Cognitive consequences of formal clothing: The effects of clothing versus thinking of clothing. *Comprehensive Results in Social Psychology*, *2*, 228–252.

Burns, D. M., Fox, E. L., Greenstein, M., Olbright, G., & Montgomery, D. (2019). An old task in new clothes: A preregistered direct replication attempt of enclothed cognition effects on Stroop performance. *Journal of Experimental Social Psychology*, *83*, 150- 156.

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, *4*.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H.

(2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science*, *21*, 1363- 1368.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115-144.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., ... & De Rosario, M. H. (2018). Package 'pwr'. *R package version*, 1(2).

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920-980.

Cicero, D. C., Martin, E. A., Becker, T. M., & Kerns, J. G. (2016). Decreased self-concept clarity in people with schizophrenia. *The Journal of Nervous and Mental Disease*, *204*(2), 142-147.

Civile, C., & Obhi, S. S. (2017). Students wearing police uniforms exhibit biased attention toward individuals wearing hoodies. *Frontiers in Psychology*, *8*, 1–14.

Cohen, J. (1988). 1988: Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Coyne, S. M., Rogers, A., Shawcroft, J., & Hurst, J. L. (2021). Dressing up with Disney and make-believe with Marvel: The impact of gendered costumes on gender typing, prosocial behavior, and perseverance during early childhood. *Sex Roles, 85*, 1-12.

Cutright, K. M., Srna, S., & Samper, A. (2019). The aesthetics we wear: How attire influences what we buy. *Journal of the Association for Consumer Research*, *4*, 387- 397.

Dubois, D., & Anik, L. (2020). From style to status and to power: When and why do stylistic choices in footwear make women feel and act powerful? *Advances in Strategic Management, 34,* 77-100.

Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: Can the "Macbeth Effect" be replicated? *Basic and Applied Social Psychology*, *36*, 91-98.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621.

Ellis, D. A., & Jenkins, R. (2015). Watch-wearing as a marker of conscientiousness. *PeerJ*, *3*, e1210–11.

Fisher, Z., & Tipton, E. (2015). Robumeta: An R-package for robust variance estimation in meta-analysis. *arXiv preprint arXiv:1503.02220*.

Fisher, Z., Tipton, E., & Hou, Z. (2017). "robumeta: Robust Variance Meta-Regression." R package Version 2.0., <https://CRAN.R-project.org/package=robumeta>.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic bulletin & review*, *21*, 1180-1187.

Frank, M. G., & Gilovich, T. (1988). The dark side of self and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology, 54*, 74–85.

Fredrickson, B. L., Roberts, T. A., Noll, S. M., Quinn, D. M., & Twenge, J. M. (1998). That swimsuit becomes you: Sex differences in self-objectification, restrained eating, and math performance. *Journal of Personality and Social Psychology, 75*, 269-284.

Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science*, *12*(6),

1077-1099.

Formann, A. K. (2008). Estimating the proportion of studies missing for meta-analysis due to publication bias. *Contemporary clinical trials*, *29*(5), 732-739.

Gamble, T., & Walker, I. (2016). Wearing a bicycle helmet can increase risk taking and sensation seeking in adults. *Psychological Science, 27*, 289-294.

Garrison, K. E., Tang, D., & Schmeichel, B. J. (2016). Embodying power: A preregistered replication and extension of the power pose effect. *Social Psychological and Personality Science*, *7*, 623-630.

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

Glenberg, A. M. (1997). What memory is for. *The Behavioral and Brain Sciences, 20*, 1–55 Hannover, B., & Kuhnen, U. (2002). "The clothing makes the self" via knowledge activation. *Journal of Applied Social Psychology, 32*, 2513-2525.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535-549.

Hebl, M. R., King, E. B., & Lin, J. (2004). The swimsuit becomes us all: Ethnicity, gender, and vulnerability to self-objectification. *Personality and Social Psychology Bulletin, 30*, 1322-1331.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39-65.

Ishii, K., Numazaki, M., & Tado'oka, Y. (2019). The effect of pink/blue clothing on implicit and explicit gender- related self- cognition and attitudes among men. *Japanese Psychological*

Research, *61*, 123-132.

Johnson, R. D., & Downing, L. L. (1979). Deindividuation and valence of cues: Effects on prosocial and antisocial behavior. *Journal of Personality and Social Psychology, 37*, 1532-2538.

Johnson, K., Lennon, S. J., & Rudd, N. (2014). Dress, body, and self: Research in the social psychology of dress. *Fashion and Textiles, 1,* 1–24.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., ... & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569-586.

Jones, M. G., Lee, T., Chesnutt, K., Carrier, S., Ennes, M., Cayton, E., Madden, L., & Huff, P. (2019). Enclothed cognition: Putting lab coats to the test. *International Journal of Science Education*, *41*, 1962-1976.

Karl, K. A., Hall, L. M., & Peluchette, J. V. (2013). City employee perceptions of the impact of dress and appearance: You are what you wear. *Public Personnel Management*, *42*, 452-470.

Kellerman, J. M., & Laird, J. D. (1982). The effect of appearance on self- perceptions. *Journal of Personality*, *50*, 296-351.

Kouchaki, M., Gino, F., & Jami, A. (2014). The burden of guilt: Heavy backpacks, light snacks, and enhanced morality. *Journal of Experimental Psychology: General*, *143*, 414-424.

Kraus, M. W., & Mendes, W. B. (2014). Sartorial symbols of social class elicit class consistent behavioral and physiological responses: A dyadic approach. *Journal of Experimental Psychology: General*, *143*, 2330-2340.

Kwon, Y. (1994). The influence of appropriateness of dress and gender on the self-perception of

occupational attributes. *Clothing and Textile Research Journal, 12*, 33-39.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 863.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389-402.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, 26(12), 1827-1832.

López-Pérez, B., Ambrona, T., Wilson, E. L., & Khalil, M. (2016). The effect of enclothed cognition on empathic responses and helping behavior. *Social Psychology*, *47*, 223- 231.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology*, 45, 216- 222.

Martins, Y., Tiggemann, M., & Kirkbride A. (2007). Those speedos become them: The role of self-objectification in gay and heterosexual men's body image. *Personality and Social Psychology Bulletin, 33*, 634-647.

Mathur, M. B., & VanderWeele, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*, *38*(8), 1336–1342.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487-498.

Mendoza, S. A., & Parks-Stamm, E. J. (2020). Embodying the police: The effects of enclothed cognition on shooting decisions. *Psychological Reports*, *123*, 2353-2371.

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., ... & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, *113*(1), 34-58.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review, 9*, 184–211.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Ogle, J. P., Tyner, K. E., & Schofield-Tomschin, S. (2013). The role of maternity dress consumption in shaping the self and identity during the liminal transition of pregnancy. *Journal of Consumer Culture*, *13*, 119-139.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531-536.

Peluchette, J. V., & Karl, K. (2007). The impact of workplace attire on employee self-perceptions. *Human Resource Development Quarterly*, *18*, 345-360.

Peña, J., Hancock, J. T., & Merola, N. A. (2009). The priming effects of avatars in virtual settings. *Communication Research*, *36*, 838-856.

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, *90*(1), 175.

Plummer, M., Stukalov, A., Denwood, M. (2022). rjags: Bayesian Graphical Models using

MCMC_. R package version 4-13, <https://CRAN.R-project.org/package=rjags>.

Rosenberg, M. S. (2010). A generalized formula for converting chi-square tests to effect sizes for meta-analysis. *PloS One*, *5*(4), e10059.

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57-71.

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425-438.

Rafaeli, A., Dutton, J., Harquail, C. V., & Mackie-Lewis, S. (1997). Navigating by attire: The use of dress by female administrative employees. *Academy of Management Journal*, *40*, 9-45.

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331-363.

Schimmack, U., & Bartoš, F. (2023). Estimating the false discovery risk of (randomized) clinical trials in medical journals based on published p-values. *arXiv preprint arXiv:2302.00774*.

Schimmack, U., & Brunner, J. (2017). Z-curve: A method for the estimating replicability based on test statistics in original studies. Retrieved from https://replicationindex.files .wordpress.com/2017/11/adv-meth-practices-draftv17-12-08.pdf

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 364-376.

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science, 28*, 687-693.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534–547.

Simonsohn, U., Nelson, L., & Simmons, J. (2015). Official user-guide to the P-curve. Retrieved from http://www .p-curve.com/guide.pdf

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*, 1146–1152.

Singer, J. E., Brush, C. A., & Lublin, S. C. (1965). Some aspects of deindividuation: Identification and conformity. *Journal of Experimental Social Psychology, 1*, 356- 378.

Slepian, M. L., Ferber, S. N., Gold, J. M., & Rutchick, A. M. (2015). The cognitive consequences of formal clothing. *Social Psychological and Personality Science*, *6*, 661-668.

Solomon, M. R., & Schopler, J. (1982). Self-consciousness and clothing. *Personality and Social Psychology Bulletin*, *8*, 508-514.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768-777.

Sterne, J. A., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*, 99-110.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85–112.

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research: A simplified methodology. *Work-Learning Research*, *1*(9).

Tiggemann, M., & Andrew, R. (2012). Clothes make a difference: The role of self objectification. *Sex Roles*, *66*, 646-654.

Van Aert, R. C., Wicherts, J. M., & Van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*(5), 713-729.

Van Stockum, C. A., & DeCaro, M. S. (2014). Enclothed cognition and controlled attention during insight problem-solving. *The Journal of Problem Solving*, *7*, 73–83.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of experimental social psychology*, *67*, 2-12.

Vickroy, S.C., Shaw, J. B., & Fisher, C. D. (1982). Effects of temperature, clothing, and task complexity on task performance and satisfaction. *Journal of Applied Psychology*, *67*, 97-102.

Wang, X., Wang, X., Lei, J., & Chao, M. C. H. (2021). The clothes that make you eat healthy: The impact of clothes style on food choice. *Journal of Business Research*, *132*, 787-799.

White, R. E., & Carlson, S. M. (2016). What would Batman do? Self-distancing improves executive function in young children. *Developmental Science*, *19*, 419-426.

White, R. E., Prager, E. O., Schaefer, C., Kross, E., Duckworth, A. L., & Carlson, S. M. (2017). The “Batman Effect”: Improving perseverance in young children. *Child Development, 88*, 1563-1571.

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, *322*, 606-607.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychological Methods*, *6*(4), 413-429.

Zhong, C. B., Bohns, V. K., & Gino, F. (2010). Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. *Psychological Science*, *21*, 311-314.

Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451-1452.

Zickfeld, J. H., & Schubert, T. W. (2016). Revisiting and extending a response latency measure of inclusion of the other in the self. *Comprehensive Results in Social Psychology*, *1*(1-3), 106-129.