Journal of Experimental Social Psychology 67 (2016) 68–82

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Many Labs 3: Evaluating participant pool quality across the academic semester via replication☆

Charles R. Ebersole [a,*], Olivia E. Atherton [b], Aimee L. Belanger [c], Hayley M. Skulborstad [c], Jill M. Allen [d], Jonathan B. Banks [e], Erica Baranski [f], Michael J. Bernstein [g], Diane B.V. Bonfiglio [h], Leanne Boucher [e], Elizabeth R. Brown [i], Nancy I. Budiman [b], Athena H. Cairo [j], Colin A. Capaldi [k], Christopher R. Chartier [h], Joanne M. Chung [b], David C. Cicero [l], Jennifer A. Coleman [j], John G. Conway [m], William E. Davis [n], Thierry Devos [o], Melody M. Fletcher [p], Komi German [b], Jon E. Grahe [q], Anthony D. Hermann [r], Joshua A. Hicks [n], Nathan Honeycutt [o], Brandon Humphrey [c], Matthew Janus [a], David J. Johnson [s], Jennifer A. Joy-Gaba [j], Hannah Juzeler [q], Ashley Keres [h], Diana Kinney [a], Jacqeline Kirshenbaum [r], Richard A. Klein [m], Richard E. Lucas [s], Christopher J.N. Lustgraaf [p], Daniel Martin [a], Madhavi Menon [e], Mitchell Metzger [h], Jaclyn M. Moloney [j], Patrick J. Morse [f], Radmila Prislin [o], Timothy Razza [e], Daniel E. Re [t], Nicholas O. Rule [t], Donald F. Sacco [p], Kyle Sauerberger [f], Emily Shrider [h], Megan Shultz [q], Courtney Siemsen [r], Karin Sobocko [k], R. Weylin Sternglanz [e], Amy Summerville [c], Konstantin O. Tskhay [t], Zack van Allen [k], Leigh Ann Vaughn [u], Ryan J. Walker [c], Ashley Weinberg [o], John Paul Wilson [t], James H. Wirth [v], Jessica Wortman [s], Brian A. Nosek [w]

[a] University of Virginia, USA
[b] University of California, Davis, USA
[c] Miami University, USA
[d] Montana State University, USA
[e] Nova Southeastern University, USA
[f] University of California, Riverside, USA
[g] The Pennsylvania State University Abington, USA
[h] Ashland University, USA
[i] University of North Florida, USA
[j] Virginia Commonwealth University, USA
[k] Carleton University, CA
[l] University of Hawaii at Manoa, USA
[m] University of Florida, USA
[n] Texas A&M University, USA
[o] San Diego State University, USA
[p] The University of Southern Mississippi, USA
[q] Pacific Lutheran University, USA

☆ *Authors' note*: This project was supported by the Center for Open Science. The authors declare no conflict of interest with the research. We thank Jessi Smith and the Montana State Motivation and Diversity Research Lab Assistants, Raelyne L. Dopko and John M. Zelenski from the Carleton University Happiness Laboratory, Idania Elizabeth Cater, Karissa Delorme, Elizabeth Friedman, Emma Hayden, Patricia Herrmann, Christine Hill, Kanako Kambe, Abigail Kane-Gerard, Valeria Lopez, Guy Merus, Andrea Moran, Alexis Murphy, Hayley Paquette, Stephanie Ponce, Candace Sowle, Boglarka Varga, and Abraham Yacaman for their assistance with data collection. We also thank Adam Galinsky, Ena Inesi, Benoît Monin, Anne Wilson, Daniel Kahneman, Michael Inzlicht, Hans IJzerman, Aleksandra Szymków-Sudziarska, Richard Petty, Lera Boroditsky, Nils Jostmann, Daniel Lakens, Robert Cialdini, and Norbert Schwarz for their suggestions and advice when designing this study. The first four and last authors comprised the project coordinating team.

* Corresponding author at: University of Virginia, 102 Gilmer Hall, PO Box 400400, Charlottesville, VA 22904, USA.
E-mail addresses: cebersole@virginia.edu (C.R. Ebersole), oeatherton@ucdavis.edu (O.E. Atherton), belangal@miamioh.edu (A.L. Belanger), skulbohm@MiamiOH.edu (H.M. Skulborstad), jill.allen1@montana.edu (J.M. Allen), jb2676@nova.edu (J.B. Banks), ericanbaranski@gmail.com (E. Baranski), mjb70@psu.edu (M.J. Bernstein), dbonfig@ashland.edu (D.B.V. Bonfiglio), leanne.boucher@nova.edu (L. Boucher), elizabeth.r.brown@unf.edu (E.R. Brown), nibudiman@ucdavis.edu (N.I. Budiman), cairoah@vcu.edu (A.H. Cairo), colin_capaldi@carleton.ca (C.A. Capaldi), cchartie@ashland.edu (C.R. Chartier), jmhchung@ucdavis.edu (J.M. Chung), dcicero@hawaii.edu (D.C. Cicero), colemanj3@vcu.edu (J.A. Coleman), john.conway@ufl.edu (J.G. Conway), dbillium@gmail.com (W.E. Davis), tdevos@mail.sdsu.edu (T. Devos), melody.fletcher@eagles.usm.edu (M.M. Fletcher), ktgerman@ucdavis.edu (K. German), graheje@plu.edu (J.E. Grahe), ahermann@bradley.edu (A.D. Hermann), joshua.hicks@gmail.com (J.A. Hicks), nhoneycutt@mail.sdsu.edu (N. Honeycutt), humphrbt@miamioh.edu (B. Humphrey), mj3at@virginia.edu (M. Janus), djjohnson@smcm.edu (D.J. Johnson), jjoygaba@vcu.edu (J.A. Joy-Gaba), juzelehj@plu.edu (H. Juzeler), akeres@ashland.edu (A. Keres), dmk2wa@virginia.edu (D. Kinney), jkirshenbaum@mail.bradley.edu (J. Kirshenbaum), raklein@ufl.edu (R.A. Klein), lucasri@msu.edu (R.E. Lucas), cjn.lustgraaf@gmail.com (C.J.N. Lustgraaf), dpmartin42@gmail.com (D. Martin), madhavi@nova.edu (M. Menon), mmetzger@ashland.edu (M. Metzger), moloneyjm@vcu.edu (J.M. Moloney), pmors001@ucr.edu (P.J. Morse), rprislin@mail.sdsu.edu (R. Prislin), razzatim@nova.edu (T. Razza), dan.re@utoronto.ca (D.E. Re), rule@psych.utoronto.ca (N.O. Rule), Donald.Sacco@usm.edu (D.F. Sacco), kylesauerberger@gmail.com (K. Sauerberger), eshrider@ashland.edu (E. Shrider), shultzmr@plu.edu (M. Shultz), csiemsen@mail.bradley.edu (C. Siemsen), KarinSobocko@cmail.carleton.ca (K. Sobocko), sterngla@nova.edu (R. Weylin Sternglanz), amy.summerville@miamioh.edu (A. Summerville), konstantin.tskhay@gmail.com (K.O. Tskhay), vanallen.22@gmail.com (Z. van Allen), Lvaughn@ithaca.edu (L.A. Vaughn), walkerrj@miamioh.edu (R.J. Walker), alweinb@gmail.com (A. Weinberg), jp.wilson@utoronto.ca (J.P. Wilson), wirth.48@osu.edu (J.H. Wirth), wortmanj@msu.edu (J. Wortman), nosek@virginia.edu (B.A. Nosek).

http://dx.doi.org/10.1016/j.jesp.2015.10.012
0022-1031/

[r] Bradley University, USA
[s] Michigan State University, USA
[t] University of Toronto, CA
[u] Ithaca College, USA
[v] The Ohio State University at Newark, USA
[w] University of Virginia, Center for Open Science, USA

ARTICLE INFO

Article history:
Received 9 March 2015
Revised 27 September 2015
Accepted 29 October 2015

Keywords:
Social psychology
Cognitive psychology
Replication
Participant pool
Individual differences
Sampling effects
Situational effects

ABSTRACT

The university participant pool is a key resource for behavioral research, and data quality is believed to vary over the course of the academic semester. This crowdsourced project examined time of semester variation in 10 known effects, 10 individual differences, and 3 data quality indicators over the course of the academic semester in 20 participant pools ($N = 2696$) and with an online sample ($N = 737$). Weak time of semester effects were observed on data quality indicators, participant sex, and a few individual differences—conscientiousness, mood, and stress. However, there was little evidence for time of semester qualifying experimental or correlational effects. The generality of this evidence is unknown because only a subset of the tested effects demonstrated evidence for the original result in the whole sample. Mean characteristics of pool samples change slightly during the semester, but these data suggest that those changes are mostly irrelevant for detecting effects.

University participant pools provide access to participants for a great deal of published behavioral research. The typical participant pool consists of undergraduates enrolled in introductory psychology courses that require students to complete some number of experiments over the course of the academic semester. Common variations might include using other courses to recruit participants or making study participation an option for extra credit rather than a pedagogical requirement. Research-intensive universities often have a highly organized participant pool with a participant management system for signing up for studies and assigning credit. Smaller or teaching-oriented institutions often have more informal participant pools that are organized ad hoc each semester or for an individual class.

To avoid selection bias based on study content, most participant pools have procedures to avoid disclosing the content or purpose of individual studies during the sign-up process. However, students are usually free to choose the time during the semester that they sign up to complete the studies. This may introduce a selection bias in which data collection on different dates occurs with different kinds of participants, or in different situational circumstances (e.g., the carefree semester beginning versus the exam-stressed semester end).

If participant characteristics differ across time during the academic semester, then the results of studies may be moderated by the time at which data collection occurs. Indeed, among behavioral researchers there are widespread intuitions, superstitions, and anecdotes about the "best" time to collect data in order to minimize error and maximize power. It is common, for example, to hear stories of an effect being obtained in the first part of the semester that then "disappears" in a follow-up study collected at the end of the semester. Beliefs about this variation can be so strong that some laboratories adopt policies to avoid data collection during particular time periods.

Are these concerns warranted? There is some evidence that individual differences among participants vary slightly across the academic semester (Table 1), but there is almost no evidence to indicate whether that variation *on average* has any impact on the detectability and effect magnitudes of correlational or experimental results. We investigated variation in detectability of 10 previously reported effects across 20 participant pools ($N = 2696$) and an online resource ($N = 737$).

1. Time of semester effects: legitimate concern or superstition?

Concerns about time-of-semester effects are not new. The existing evidence supports the belief that participants at the beginning of the semester are different on average from participants at the end of the semester. However, the differences are modest. For example, later participation in the semester is related to lower levels of conscientiousness (Witt, Donnellan, & Orlando, 2011) and higher levels of openness to experience (Aviv, Zelenski, Rallo, & Larsen, 2002; see Table 1). In addition, individuals who participate late in the semester show lower intrinsic motivation when compared to those who participated earlier (Hom, 1987; Nicholls, Loveless, Thomas, Loetscher, & Churches, 2015).

Research on variation in actual task performance, however, has produced mixed results. For instance, Wang and Jentsch (1998; $N = 49$) asked participants to complete a cued recall task, testing their memory for the English meanings of 24 learned foreign words after a 30-min period. They found no significant difference in cued recall between the earliest and latest participants over the course of four semesters.

In contrast, Nicholls et al. (2015) did find evidence for differential sustained attention across the semester. In their study ($N = 80$), individuals who participated either for course credit or monetary compensation completed hundreds of trials of a reaction time-based number detection task (Sustained Attention to Response Task; Robertson, Manly, Andrade, Baddeley, & Yiend, 1997) at either the beginning or end of the semester. There were no significant differences between course credit participants and paid participants at the beginning of the semester. However, paid participants outperformed course credit participants at the end of the semester, $F(1, 37) = 5.58$, $p = .024$, $\eta_p^2 = .131$, possibly related to the latter group's relatively lower levels of intrinsic motivation.

2. Research questions

The present project is informally called "Many Labs 3" as it follows the model established in two prior investigations for conducting the identical procedure in many different laboratories (Klein et al., 2014, 2015). In

Table 1
Correlations between time of semester and Big Five personality traits.

	Aviv, Zelenski, Rallo, & Larsen (2002; using NEO-PI R; $N = 257$)	Witt et al. (2011); using IPIP-NEO; $N = 512$)
Agreeableness	−.11	−.10
Conscientiousness	−.14	−.20
Extraversion	.19	.02
Neuroticism	−.11	−.08
Openness	.14	−.01

Note: Values represent Pearson's *r* between personality trait and week of participation.

Many Labs 3, we investigated the extent to which 10 psychological effects and multiple individual difference variables varied across the academic semester. The same experimental procedure was administered in 20 participant pools at institutions in the United States and Canada. This allowed us to investigate the extent to which participant characteristics and the magnitudes of different effects vary across the academic semester. If time of semester effects were observed, we also obtained a Mechanical Turk sample (MTurk; $N = 737$) to help distinguish between time of semester effects (unique to students) versus time of year effects.

A secondary interest was to provide additional evidence about the included effects using large scale replication: their overall effect size, variation by site and sample, and moderation by time of semester. Some of the effects we included are heavily studied, but others are relatively new or have not been replicated frequently enough to clarify boundary conditions or moderating influences. The final materials and dataset will be of substantial use beyond this initial report, particularly to explore moderating influences not examined for this report. All data and materials are available for additional investigation by others (https://osf.io/ct89g/).

3. Method

3.1. Participants

An open invitation for researchers to participate as a data collection site was issued in early 2014 for data collection to occur from August through December. To be eligible for inclusion, participating labs agreed to administer the study procedure to *at least* 80 participants total with at least 40 from the first half of the semester and at least 40 from the second half of the semester. To ensure that teams were operating on similar academic calendars, participation was limited to institutions in the United States and Canada.

Twenty teams completed the data collection with the average sample size being 135.40 ($SD = 63.00$), ranging from 45 to 321 (see Table S1 for details of each team and Table S2 for characteristics of each participant pool). One team was unable to meet the minimum participant cutoff ($N = 45$), but earned authorship through other contributions. Their data are included in the aggregate set and all subsequent analyses. Overall, 69.8% of the sample were female, the average age was 19.3 years ($SD = 3.7$), and 53.7% were White, 9.4% Black, 16.0% Asian, 10.6% Hispanic, and 10.3% other.

These participants came from a wide range of institutions, producing a relatively diverse undergraduate sample. Although all of the directly replicated effects collected data from undergraduate participants, the current sample differs in a few ways. None of the original study collection sites are represented in the current sample. Two original studies recruited undergraduates independent of a participant pool, and two other original studies were conducted at European institutions. Finally, the current sample has a heavier representation of females compared to original studies that reported this demographic (55.5%). Sample differences that seem particularly relevant are noted in the descriptions of each effect.

We simultaneously collected participants from MTurk over the same time period ($N = 737$) as a comparison sample for time of year effects and sample diversity. In the MTurk sample, 48.6% of the sample were female, the average age was 35.1 years ($SD = 10.9$), and 66.4% were White, 15.4% Asian, 7% Black, 4.7% Hispanic, and 6.5% other. This sample was drawn from the United States and there were no requirements for previous MTurk experience (e.g., minimum number of previous HITs completed). MTurk participants received $1.25 as compensation for their time.

3.2. Selection of effects

The primary aim of the project was to detect possible variability in effect magnitudes across the academic semester when using university participant pools. To obtain a candidate list of effects and individual difference measures, we held a round of open nominations and invited submissions for any effect that fit the defined criteria. Those nominations were supplemented by ideas from the project team and from direct queries to independent experts in psychological science. Given the areas of interest of the project coordinators and most collaborators, nominations came largely from the fields of social and personality psychology.

The coordinating team sought effects and individual difference measures that fit the following criteria: (1) highly feasible implementation through a web browser or in the lab, (2) brevity of study procedures, and (3) high interest value of the theoretical domain or phenomenon. In addition, for the collected set of effects and measures we sought: (1) diversity of represented research domains, (2) diversity of known or presumed likelihood of variation across the semester, and (3) diversity of "classic" well-established effects and contemporary effects that have untested replicability.

The project coordinating team collectively evaluated the nominated studies (see Table S3 for a list of considered effects). No specific researcher was "targeted" for replication because of concerns or skepticism about an effect. In fact, any included effect that was not reproducible at all would produce little insight about variation across the semester, which was the central research question for this project. Given this, one strategy would have been to only select classic, well-established effects for replication. However, it is possible that these effects are well established *because* they are resistant to contextual variation. Had we selected only well-established effects, we could have undermined the possibility of observing context effects. Our presumption was that time-of-semester effects are most likely to occur for so-called "fragile" effects that might be particularly sensitive to context. As such, we included high-profile, contemporary effects with less certain replicability, particularly from domains in which popular debate suggests fragility or sensitivity to context.

This project was most concerned with detecting whether or not time of semester variation happens in regular research practices. Therefore, if we had limited our effects to one or two research domains (e.g., effects moderated by attention, Nicholls et al., 2015), we might have maximized testing "*can* semester variation alter effects?", but sacrificed testing "*does* time of semester variation alter effects?" in ordinary research practice. Furthermore, reduced attention can be reasonably hypothesized as moderators for many effects, even if they have not been previously demonstrated as influential. In other words, we aimed to examine time-of-semester as the highly available explanation when two behavioral lab studies show different results, whatever the topic of study.

Once selected for inclusion, a member of the research team contacted the corresponding author (if alive) to obtain the original study materials and get advice about adapting the procedure for use in this study.[1] In particular, we asked the original authors if there were moderators or other limitations to obtaining the result that would be useful for the team to understand in advance or to anticipate during data collection. The team implemented a draft of the proposed study procedure and solicited feedback from the original authors to further improve the design. This process was undertaken to minimize reasons to expect different outcomes between the original outcomes and the replications. Sometimes this led to adaptations of the procedure in order to maximize its relevance in the present context, or changes to fit the constraints of the present procedure (see Table S4 for a summary of procedure adaptations). Also, some initially selected effects were eliminated during review if we could not address a priori design concerns effectively.

We implemented a draft study procedure to pre-test for length. Data collection constraints required completion of all study materials within 30 min. A pilot sample of 30 volunteers completed the on-line portion of

[1] In the case of a conceptual replication of the relationship between persistence and conscientiousness, we did not follow this procedure and seek original materials.

the study procedure. We calculated the time required for 85% of participants to complete each study procedure. Following this piloting, we needed to remove three individual difference measures, shorten one procedure (Stroop task), and eliminate two effects to meet the time constraints. After this intensive review, 10 effects, 10 individual difference measures, 3 data quality indicators, and a selection of demographics items were confirmed for inclusion in Many Labs 3. In administration of the actual procedure, we did not impose a 30 min time constraint, but individual data collection sites could let participants go before data collection completion if circumstances demanded it. 97.2% of non-MTurk participants completed the entire study.

3.3. Procedure

The study procedures and materials were reviewed and approved by the University of Virginia Institutional Review Board for the Social and Behavioral Sciences as well as IRBs from all other participating institutions.

Eight of the effects were administered in a single computerized experiment script that began with informed consent, then presented the procedures for each target effect in a random order, then presented the ten individual difference measures and three data quality indicators, and closed with demographics items and debriefing. Two of the effects could not be administered via computer, one because the participants were required to hold the measures in their hands (Weight Embodiment) and another because the original author suggested that it required a paper–pencil administration format (Metaphoric Restructuring). As such, the participant was instructed to go to the experimenter for instructions at a random point during presentation of the eight computerized tasks. At this point, the two "in-person" tasks were administered in a counterbalanced order. The script for the experiment and video simulations of experiment administration are available publicly (https://osf.io/ct89g/).

The procedure for the MTurk sample was the same except that we removed the two "in-person" tasks and one of the computer-administered tasks that involved deception and concerned an issue at the participant's university (Elaboration Likelihood).

3.4. Demographics measures

3.4.1. Age

Participants noted their age in years in an open-response box.

3.4.2. Sex

Participants selected "male" or "female" to indicate their biological sex.

3.4.3. Race/ethnicity

Participants from sites in the United States indicated their race/ethnicity by selecting: African–American, Asian–American, Native Hawaiian and other Pacific Islander, Latino or Hispanic, Native American and Alaska Native, White, non-Hispanic or Latino, or Multiracial. Those in Canada selected from: Caucasian, White; Black (African, African American); South Asian, Indian, Pakistani, etc.; East Asian, Chinese, Japanese, etc.; Arabic, Central Asian; Hispanic, Central, or South American; Aboriginal. Participants at all sites could also select "Other" and write a response.

3.4.4. Year in college

Participants responded to an item, "What year in college are you?" by indicating whether they are a: Freshman (first-year), Sophomore (second-year), Junior (third-year), Senior (fourth-year). Participants could also select "Other" and write a response.

3.4.5. College major

Participants indicated their major in an open-response box.

3.4.6. Data quality indicators

Several items at the end of the study, just prior to the demographics items, assessed carelessness or lack of effort.

3.4.6.1. Participation questions. To assess the quality of the participant's engagement in the study, we asked: "How much effort did you put into the tasks during this experiment?" (1 = *no effort* to 5 = *I tried my hardest*) and "How closely did you pay attention to the instructions and tasks during the experiment?" (1 = *none* to 5 = *I gave the tasks my undivided attention*).

Participants also responded to items assessing: (1) whether they were participating as part of a class requirement, extra credit, payment, or other; (2) the type of class that required/incentivized this participation (i.e., introductory course in psychology, secondary/upper-division course in psychology, any class above secondary, research methods/statistics course, or other); and, (3) if required, how close they were to completing their subject pool requirements (this is my first study, about 25% done, about 50% done, about 75% done, this is my last study, I am not participating for a class requirement).

3.4.6.2. Instructional attention check. The instructional attention check presented a paragraph of instructions in which the last sentence read: "So, in order to demonstrate that you have read the instructions, please ignore the preferences form below, and simply write 'I read the instructions' in the box below." Immediately below this paragraph is an item saying "In my free time I prefer:" with response options of (1) engaging in hobbies, (2) watching TV, reading, music, (3) being in nature, (4) exercising, (5) cooking or eating, and (6) other (with an open response area for writing in the correct answer).

3.5. Individual difference measures

Brief individual difference measures were selected as possible moderators of psychological effects based on prior evidence that participant characteristics vary across the semester or because of their widespread use in psychological science. Table 2 shows the descriptive statistics for all of the individual differences measures (see Table S5 for correlations among these measures). When comparisons were available, reliabilities for measures were similar to or better than prior uses.

3.5.1. Global self-esteem (Robins, Hendin, & Trzesniewski, 2001)

Global self-esteem was measured using a Single-Item Self-Esteem Scale (SISE) designed as an alternative to using the Rosenberg Self-Esteem Scale (1965). The SISE consists of a single item: "I have high self-esteem." Participants respond on a 7-point Likert scale, ranging from 1 (not very true of me) to 7 (very true of me). Robins et al. (2001) reported strong convergent validity with the Rosenberg Self-Esteem Scale (with *r*s ranging from .70 to .80) among adults. Further, the item had similar predictive validity to the Rosenberg Self-Esteem Scale.

Table 2
Descriptive statistics of individual differences measures.

	M (*SD*)	α	Scale range
Conscientiousness	5.47 (1.20)	.52	1–7
Agreeableness	4.95 (1.17)	.36	1–7
Neuroticism	3.40 (1.42)	.67	1–7
Openness to experience	5.18 (1.14)	.41	1–7
Extraversion	4.18 (1.59)	.72	1–7
Intrinsic motivation	2.83 (.41)	.78	1–4
Perceived stress	2.68 (.74)	.67	1–5
Mood	5.11 (1.18)	.91	1–7
Self-esteem	4.78 (1.59)	N/A	1–7
Effort	3.83 (.82)	N/A	1–5
Attention	4.06 (.76)	N/A	1–5
Need for cognition	3.22 (.63)	.67	1–5

3.5.2. Ten-item personality inventory for big-five personality (Gosling, Rentfrow, & Swann, 2003)

We measured five dimensions of human personality (Goldberg, 1981)—conscientiousness, agreeableness, neuroticism/emotional stability, openness/intellect, and extraversion—with the Ten Item Personality Inventory (TIPI; Gosling et al., 2003). Each trait was assessed with two items on 7-point response scales from 1 (disagree strongly) to 7 (agree strongly). Reliabilities are somewhat lower than other, longer scales, but the five scales show satisfactory retest reliabilities (cf. Gnambs, 2014) and substantial convergent validities with longer Big Five instruments (e.g., Ehrhart et al., 2009; Gosling et al., 2003; Rojas & Widiger, 2014).

3.5.3. Daily mood (adapted from Schwarz and Clore (1983))

We measured daily mood using two items that assess the extent to which the participant is in a good or bad mood. Items begin with the same statement, "Today I generally feel..." Each set of response options is on a 7-point Likert scale, ranging from 1 (very unhappy) to 7 (very happy), and 1 (very bad) to 7 (very good).

3.5.4. Perceived Stress Scale—short form (Cohen, Kamarck, & Mermelstein, 1983)

We measured perceived stress over the last week using a 4-item short-form scale that is an alternative to the original, 14-item Perceived Stress Scale (Cohen et al., 1983). Participants respond on a 5-point Likert scale, ranging from 0 (never) to 4 (very often). The original study suggested that the shortened scale was relatively reliable ($\alpha = .72$) and the factor structure was consistent with the long form.

3.5.5. Need for Cognition Scale (adapted from Cacioppo and Petty (1982); Skulborstad, unpublished data)

We measured need for cognition with six items that ask about the degree to which the participant enjoys engaging in complex, deliberative, and abstract thinking. Each of the items are on a 5-point Likert scale, ranging from 1 (extremely uncharacteristic) to 5 (extremely characteristic). Following past research we selected the top six factor loading items of the original scale (e.g., Verplanken, 1991; Verplanken, Hazenberg, & Palenewen, 1992; Skulborstad, unpublished data). We used this shortened version instead of the 34 item (Cacioppo & Petty,

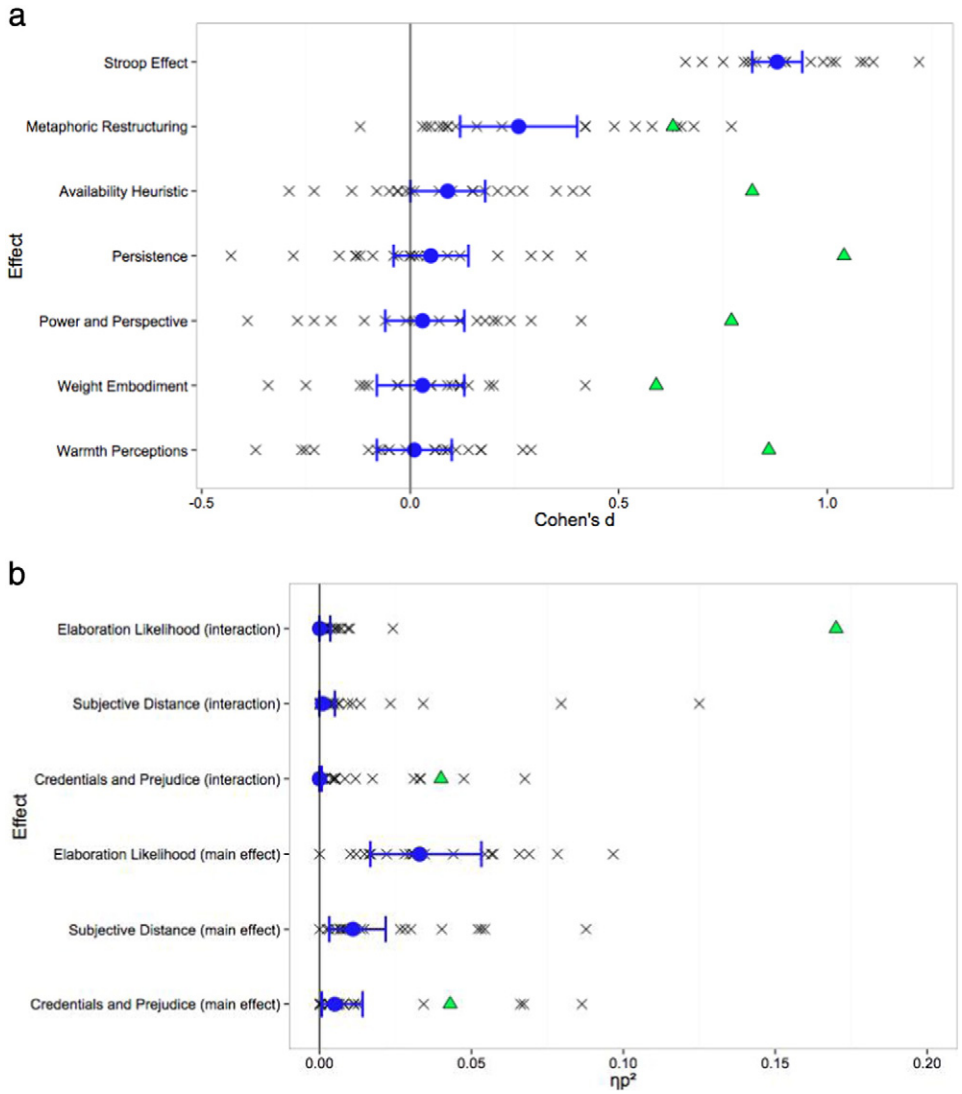

Fig. 1. Replication results organized by replication effect size, 1a for Cohen's *d* estimates, 1b for η_p^2 estimates. When available, the triangle indicates the effect size obtained in the original study (Stroop effect and Elaboration Likelihood main effect estimate do not appear because they were very large, $d = 2.04$ and $\eta_p^2 = .59$ respectively). Large circles represent the aggregate effect size obtained across all participants. Error bars represent 99% noncentral confidence intervals around the effects. Small x's represent the effect sizes obtained within each site.

1982) or 18 item versions (Petty, Cacioppo, & Kao, 1984) because of time constraints.

3.5.6. Work preference inventory, intrinsic motivation (Amabile, Hill, Hennessey, & Tighe, 1994)

We used the 15-item general intrinsic motivation scale of the Work Preference Inventory to measure the extent to which the participant is motivated because the work itself is satisfying or intriguing. The items are rated on a 4-point Likert scale, ranging from 1 (Never or almost never true of me) to 4 (Always or almost always true of me). This scale is convergent with other forms of measured motivation, but also discriminable from measures of social desirability and intelligence.

4. The effects

Next, we describe the 10 selected effects with an abstract reporting the main idea of the original research with the sample size, inferential test, and effect size. Details on the methodology and analysis plan that was defined in the pre-registered protocol for each effect can be found presented in the supplementary material (https://osf.io/ct89g/). We report the aggregate result of the replications at the end of each subsection; these results are summarized in Fig. 1a, b, and Table 3.

The focus of this replication project is to estimate the variability in effect magnitude by time of semester. As such, we aimed to identify or simplify original study designs that could be tested as two-condition experiments or as correlations when possible. Some original studies had additional conditions that were relevant for the theoretical purposes of the investigation. In those cases, the replication designs identified the key conditions relevant for estimating the effect. Also, in some cases, multiple dependent variables were included in the original design. If the dependent variables could be administered quickly, they were usually retained in the replication. When multiple outcomes were included, because they are likely to be correlated, just one or an aggregate was identified as the primary object for replication and examining variation across the semester; the others were considered secondary. Secondary outcome measures are reported in footnotes or the supplemental material. Finally, correspondence with original authors during the design process identified some potential moderating influences that could be examined with additional analyses.

4.1. Stroop task (Stroop, 1935)

In the Stroop task (Stroop, 1935), participants view words one at a time in different colors. Participants categorize the color of the font and do not need to do anything with the meaning of the word. This task is more difficult when there is a discrepancy between the color of the font and the word. For example, it is easier to categorize the font as "blue" when it is presented on the word "tree" or the word "blue" compared to being presented on the word "red." The meaning of the word "red" interferes with categorization of the font color as "blue." This task is very robust and has been used in thousands of research applications (MacLeod, 1991). Effects on the Stroop task can be larger when participants are tired, or otherwise cognitively or emotionally depleted, because they have fewer available resources to overcome the response competition. In the present study, we incorporated a simple version of the Stroop task to test whether similar variation would be observed across the semester cycle.

The Stroop task is a within-person experiment with two response conditions – font color *congruent* with color word and font color *incongruent* with color word – and response latency as a dependent variable. We used the *D* scoring algorithm for analysis of these data (Greenwald, Nosek, & Banaji, 2003), an analysis technique that has general application to response latency contrasts (Nosek & Sriram,

Table 3
Original and replication results.

Effect	ES stat	Original study			Meta-analytic estimate			Aggregate estimate			Null hypothesis significance tests by sample			Null hypothesis significance tests of aggregate			
		ES	95% CI lower, upper	Median replication ES	Replication ES	95% CI lower, upper		Replication ES	95% CI lower, upper		Proportion < 0 ($p < .05$)	Proportion > 0 ($p < .05$)	*Proportion ns*	Key statistics	df	N	p
Stroop task	*d*	2.04	1.90, 2.18	0.89	0.91	0.84, 0.98		0.88	0.84, 0.92		0.00	1.00	0.00	t = 49.795	3336	3337	<.001
Metaphoric restructuring	*d*	0.63	0.07, 1.20	0.19	0.29	0.17, 0.42		0.26	0.15, 0.37		0.00	0.25	0.75	$\chi 2 = 21.90$	1	1335	<.001
Availability heuristic	*d*	0.82	0.47, 1.17	0.07	0.09	0.02, 0.16		0.09	0.02, 0.16		0.00	0.14	0.86	Psdep = .522	N/A	3088	0.015
Power and perspective	*d*	0.77	0.12, 1.41	0.03	0.03	−0.04, 0.11		0.03	−0.04, 0.10		0.00	0.05	0.95	t = .89	2967	2969	0.37
Weight embodiment	*d*	0.59	0.01, 1.16	0.05	0.03	−0.05, 0.11		0.03	−0.06, 0.11		0.00	0.00	1.00	t = .61	2283	2285	0.543
Warmth perceptions	*d*	0.86	0.40, 1.33	0.06	0.01	−0.08, 0.06		0.01	−0.06, 0.08		0.05	0.00	0.95	t = .22	3117	3119	0.827
Elaboration likelihood	ηp^2	0.17	0.06, 0.29	0.000001	0.000001	0.000, 0.001		0.00005	0.000, 0.002		0.00	0.00	1.00	F = .129	1, 2361	2365	0.72
Self-esteem and subjective distance	ηp^2	n/a	n/a	0.0001	0.0004	0.000, 0.005		0.001	0.000, 0.004		0.05	0.10	0.85	F = 1.98	1, 3131	3136	0.16
Credentials and prejudice	ηp^2	0.04	0, 0.09	0.00	0.0003	0.000, 0.003		0.000	0.000, 0.000		0.00	0.00	0.90	F = .0004	1, 3130	3134	0.985
Conceptual replication																	
Persistence and conscientiousness	*d*	1.04	0.64, 1.43	0.000	0.03	−0.04, 0.10		0.05	−0.02, 0.12		0.00	0.00	1.00	r = .027	3191	3193	0.134
Added effects																	
*Elaboration likelihood – Main effect	ηp^2	0.59	0.47, 0.67	0.031	0.034	0.022, 0.049		0.033	0.020, 0.048		0.00	0.45	0.55	F = 79.925	1, 2361	2365	<.001
*Self-esteem and subjective distance – Main effect	ηp^2	n/a	n/a	0.011	0.012	0.004, 0.024		0.011	0.005, 0.019		0.00	0.24	0.76	F = 7.97	1, 3131	3136	<.001
*Credentials and prejudice – Main effect	ηp^2	0.043	0.002, 0.103	0.005	0.005	0.001, 0.014		0.005	0.001, 0.012		0.00	0.14	0.86	F = 17.01	1, 3130	3134	<.001

Note. Weighted statistics are computed on the whole aggregated dataset; Meta-analytic statistics are computed on the disaggregated dataset (N = 20 or 21). 95% CI's for original effect sizes used cell sample sizes when available and assumed equal distribution across conditions when not available. Confidence intervals around the meta-analytic mean are based on the central normal distribution. Confidence intervals around the weighted effect size are based on non-central distributions. The Stroop effect size is taken from a meta-analysis comparing Stroop performance of young adults to older adults (Verhaeghen & De Meersman, 1998). The aggregate effect size for younger adults is reported here. *For three experiments, reliable main effects were added after observing the aggregate outcomes to have more effects to then test for variation across the academic semester. Credentials and prejudice interaction effect size was estimated as 1.28 e-6, the weighted upper bound of the 95% CI was too small to compute with the statistical software. ηp^2's were not available for the original self-esteem and subjective distance effects. The Cohen's d estimates are 0.21 (95% CI 0.001, 0.418) and 0.39 (95% CI 0.18, 0.60) respectively.

2007) and avoids confounding influences in response latency comparisons that influence other analytic techniques (Sriram, Greenwald, & Nosek, 2010). First, all trials with latencies above 10,000 ms were removed. Then, we calculated the average response time for all *correct* responses separately for congruent and incongruent trials. We replaced response latencies for trials with errors using the mean of correct responses in that condition plus 600 ms. Then, we recomputed the means for congruent and incongruent trials overall. *D* is the difference between these two means divided by the standard deviation of all correct trials regardless of condition. Positive scores indicate slower response times on average for incongruent compared to congruent trials.

Across the replication studies ($N = 3{,}279$), participants took longer to categorize incongruent words than congruent words ($M = 0.27$, $SD = 0.31$), $t(3{,}336) = 50.22$, $p < .001$, $d = .88$, 95% CI = [.84, .92]. This replicated the basic effect from previous research.[2]

4.2. Metaphoric structuring: understanding time through spatial metaphors (Boroditsky, 2000, study 1)

Boroditsky (2000) demonstrated that priming participants with an ego-moving or object-moving frame of reference can influence their interpretation of an ambiguous temporal statement. Participants were given a two-page questionnaire. The first page contained four scenarios consisting of a picture and a sentence. Participants in the ego-moving condition saw scenarios describing the location of an object in reference to a stick figure (referred to as "you"). Participants in the object-moving condition saw scenarios in which two objects were described in relation to one another. Participants indicated whether the statement about the picture was true or false. On the second page, participants read an ambiguous temporal statement (e.g., "Next Wednesday's meeting has been moved forward two days") and indicated to which day the meeting had been rescheduled (e.g., "Monday" or "Friday") and how confident they felt about their choice from 1 (not at all confident) to 5 (very confident). As predicted, ego-priming was more likely to induce the answer of Friday (73.3%) than Monday (26.7%), whereas object-priming was more likely to induce the answer of Monday (69.2%) than Friday (30.8%). Overall, 71.3% of participants responded in a prime consistent manner, $\chi^2(1, N = 56) = 5.2$, $p < .05$, $d = .63$, 95% CI = [.07, 1.20]. In the control condition, 54.3% of participants selected Friday and 45.7% selected Monday.

Based on the original author's recommendations, this task was completed on paper-and-pencil in the face-to-face portion of the study to ensure comparability to the original procedure, and three conditions were included: ego-prime, object-prime, and control. We excluded participants from the analyses if, in the priming condition, they failed to answer all four priming questions (see materials) correctly, or if, in any condition, they failed to select one of the two possible correct options for the day of the meeting (Monday or Friday). In the in-lab replication studies ($N = 2{,}191$), ego-priming was more likely to induce the answer of Friday (67.8%) than Monday (32.2%), whereas object-priming showed a bias in the same direction but to a lesser extent with Friday (59.5%) being more popular than Monday (40.5%). Overall, 56.4% of participants responded in a prime consistent manner, $\chi^2(1, N = 1{,}335) = 21.90$, $p < .001$, $d = .26$, 95% CI = [.15, .37]. The effect size was weaker, and the condition differences were shifted toward selecting Friday, compared to the original study, but the replications were nonetheless consistent with the key feature of the original demonstration: object-priming increased the likelihood of selecting Monday compared to Friday. Moreover, in the control condition ($N = 856$), 63.3% of participants selected Friday and 36.7% selected Monday illustrating the overall bias toward Friday (see also Lai and Boroditsky (2013)).

4.3. Availability: a heuristic for judging frequency and probability (Tversky & Kahneman, 1973, study 3)

Tversky and Kahneman (1973) examined whether undergraduates recruited separately from a participant pool would overestimate the frequency of easier-to-imagine words relative to harder-to-imagine words. People find it easier to think of English words that begin with a certain letter (k, l, n, r, or v) than to think of words with this letter in the third position. However, these letters actually show up about twice as often in the third position compared to the first position. Participants judged whether each of these letters was more likely to show up in the first or the third position and estimated the ratio of the frequency with which they appear in each position.

Tversky and Kahneman (1973) found that 105/152 participants judged the first position to be more frequent for the majority of letters and that 47/152 participants judged the third position to be more frequent for the majority of letters. The authors reported that a sign test (Grissom & Kim, 2012) showed a significant bias favoring the first position, $PS_{dep} = 0.67$, $p = .000004$, $d = .82$, 95% CI = [.47, 1.17]. Additionally, the majority of participants judged each of the five letters to be more frequent in the first position, and the median estimated ratio for each of the five letters was 2:1.

Across the replication studies ($N = 3088$), 1612/3088 participants judged the first position to be more frequent for the majority of letters and 1476/3088 judged the third position to be more frequent for the majority of letters. The probability of favoring the first position was weak but reliable, $PS_{dep} = 0.522$, $p = .015$, $d = .09$, 95% CI = [.02, .16].[3]

In addition, participants estimated the number of times a letter appeared in the first position for every ten times it appeared in the third position. In an attempt to normalize these estimates around the point of 0 (indicating that letters occurred in both positions equally) we subjected the ratio estimate for each participant to the following transformation:

If Average Ratio = 10, then Score = 0; If Average Ratio > 10, then Score = (Average Ratio/10) − 1; If Average Ratio < 10, then Score = 1 − (10/Average Ratio). Negative scores indicate that the letters were judged to appear more frequently in the first position. To create an equal boundary for estimates above and below 10, we used only those with an average estimate greater than or equal to 1 and less than 100 (eliminating 18 of 2920 participants). Using this approach, participants estimated that the letters appeared more frequently in the first compared to the third position on average ($M = -0.79$, $SD = 1.59$), $t(2901) = -26.77$, $p < .001$, $d = -.50$, 95% CI = [−.54, −.46], and the effect size was much stronger than with the original estimation

[2] We also compared the error rates on congruent compared to incongruent trials. We calculated a per trial error rate (number of errors divided by number of trials) for each participant for congruent and incongruent trials separately. We then calculated a difference score by subtracting the congruent error rate from the incongruent error rate. Compared to a score of zero (which would indicate equal error rates), participants made more errors on incongruent trials compared to congruent trials ($M = .016$, $SD = .049$), $t(3337) = 18.52$, $d = .32$, 95% CI = [.29, .36]. Overall, participants made an average of 2.25 errors ($SD = 4.52$). As a secondary analysis, we analyzed the Stroop effect using the mean difference of log-transformed data. This alternate strategy revealed the same effect with a slightly weaker estimate, $t(3347) = 44.17$, $d = .76$, 95% CI = [.72, .80].

[3] In an exploratory analysis by letter, we observed that the first position was favored for letters K, $\chi^2(1, N = 3{,}225) = 106.84$, $p < .001$, $\Phi = .18$, and L, $\chi^2(1, N = 3242) = 49.85$, $p < .001$, $\Phi = .12$, but the third position was favored for letter N, $\chi^2(1, N = 3236) = 33.65$, $p < .001$, $\Phi = .10$, and there was no difference for letters R, $\chi^2(1, N = 3239) = 1.15$, $p = .284$, $\Phi = .02$, and V, $\chi^2(1, N = 3{,}241) = 0.682$, $p = .409$, $\Phi = .01$.

strategy (note that Fig. 1a shows data for the original estimation strategy).[4]

4.4. The relation between persistence and conscientiousness (De Fruyt, van De Wiele, & van Heeringen, 2000)

De Fruyt and colleagues (2000) investigated the relation between Cloninger's Temperament and Character Dimensions (Cloninger, 1987) and the Big Five personality index. The researchers found that Cloninger-assessed persistence correlated with the Big Five trait conscientiousness, $r(128) = .46$, $p < .001$, $d = 1.04$, 95% CI = [.64, 1.43].

The current study examined variation in persistence across the semester. To conceptually replicate the relation between persistence and conscientiousness, we used the unsolvable anagram task, which has been used as a measure of persistence (e.g., Aspinwall & Richter, 1999; Sommer & Baumeister, 2002). In this task, participants are presented with a number of anagrams to unscramble. Some anagrams are solvable, others are not. Participants choose to stop working on the task whenever they would like. Persistence is the amount of time spent on the task before moving on to the next task.

Unlike the others, this is not a direct replication. The original work examined the correlation between self-perception of persistence and a long-form personality measure using a clinical sample. We added this effect as a conceptual replication because persistence and conscientiousness are two factors frequently implicated in research and beliefs regarding time of semester variation (e.g., Aviv et al., 2002; Witt et al., 2011). No known study has examined the relationship between conscientiousness and this brief behavioral task. Moving from two self-report measures to one self-report and one behavioral measure seemed likely to reduce the estimated correlation between these constructs.

Across all replication studies ($N = 3{,}193$), there was little evidence for a relationship between conscientiousness measured with the TIPI and persistence measured with the unsolvable anagram task, $r(3191) = .027$, $p = .134$, 95% CI = [−0.008, 0.061], $d = .05$, 95% CI = [−.02, .12].

4.5. Power and perspectives not taken (Galinsky, Magee, Inesi, & Gruenfeld, 2006, study 2a)

Galinsky et al. (2006) examined whether power can impair perspective taking. Their Study 2a tested whether individuals made to feel high in power were more likely to inaccurately assume that others perceive the world from the same perspective as they do compared to those made to feel low in power. Forty-two undergraduate students wrote about a time when they had power over others (high-power condition) or about a time when someone else had power over them (low-power condition). After completing two filler tasks, they read a scenario in which they and a colleague went to a fancy restaurant. The restaurant had been recommended by the colleague's friend but they and their colleague ended up having a poor dining experience. The scenario then described their colleague sending an email to their friend who recommended the restaurant saying, "About the restaurant, it was marvelous, just marvelous." Thus, the participant knew that the response was sarcastic but the friend did not. Participants in the high-power condition thought that the colleague's friend would interpret the message as being more sarcastic and less sincere ($M = 3.74$, $SD = 1.54$) than participants in the low-power condition ($M = 4.84$, $SD = 1.30$), $t(40) = 2.47$, $p = .02$, $d = 0.77$, 95% CI = [.12, 1.41].

In an aggregate analysis of the replication studies ($N = 2{,}969$), participants in the high-power ($M = 3.75$, $SD = 1.55$) and low-power ($M = 3.80$, $SD = 1.57$) conditions thought that the sincerity of the message would be interpreted similarly, $t(2967) = 0.89$, $p = 0.37$, $d = 0.03$, 95% CI = [−.04, .10]. In sum, the replications did not provide evidence of an effect of power on perspective taking.[5]

4.6. Weight as an embodiment of importance (Jostmann, Lakens, & Schubert, 2009, study 2)

Weight is often used metaphorically to convey importance. Jostmann et al. (2009) examined whether holding heavier objects influenced participants' perceptions of importance. In their Study 2, 51 Dutch university participants (28 in the heavy clipboard condition; 23 in the light clipboard condition) stood and completed a questionnaire on a heavy (2.25 lb) or light (1.45 lb) clipboard. Participants read a scenario in which students were not allowed to express their opinion to a university committee about the size of a study abroad grant. Participants then indicated whether or not they believed that it was important for the committee to listen to the students' opinions about the grant on a scale ranging from 1 (not at all) to 7 (very much). Participants in the heavy clipboard condition ($M = 5.27$, $SD = 1.28$) believed that it was more important for the university committee to listen to the students' opinions as compared to participants in the light clipboard condition ($M = 4.21$, $SD = 2.10$), $t(49) = 2.08$, $p = .043$, $d = .59$, 95% CI = [.01, 1.16].[6]

In the replication studies, we excluded participants from the analyses if the experimenter noted any behavior that would have diffused the weight of the clipboard (e.g., sitting down, resting the clipboard on a table). Across all in-lab replications ($N = 2{,}285$), participants in the heavy ($M = 6.16$, $SD = 1.02$) and light ($M = 6.14$, $SD = 1.03$) clipboard conditions believed that it was similarly important for the university committee to listen to the students' opinions, $t(2283) = 0.61$, $p = .543$, $d = .03$, 95% CI = [−.06, .11]. Notably, the means in both conditions were within a point of the scale ceiling. Overlapping distribution plots for both conditions are presented in Fig. 2. They show very similar distributions for the two conditions, and also a bias toward selecting the top two scale points.[7]

4.7. Warmer hearts, warmer rooms (Szymkow, Chandler, IJzerman, Parzuchowski, & Wojciszke, 2013, study 1)

Many cultures use heat-based metaphors to describe personality, with communal traits often being described as "warm." Szymkow and colleagues (2013) investigated the influence of priming these metaphorically based traits on perceptions of ambient room temperature. Participants read a description of an individual who displayed either communal or agentic traits. Afterwards, participants gave their perceptions of various physical elements of the room, including an estimate of the ambient temperature. Polish university participants estimated that the ambient room temperature (°F) was warmer following the communal description ($M = 69.71$, $SD = 4.03$) compared to the agentic

[4] Participants saw the five letters in this task (K, L, N, R, V) in random order. We were interested in any effect that the progression of the task could have had on participants' estimates as to whether a given letter appeared more frequently in the first or third position. We ran analyses on each trial position (first letter seen to last letter seen) to assess any order effect within this effect. The availability bias was reliably observed on the first ($\chi^2 = 6.96$, $p = .008$, $d = .09$, 95% CI = [.02, .16]), third ($\chi^2 = 6.78$, $p = .009$, $d = .09$, 95% CI = [.02, .16]), and fourth ($\chi^2 = 7.53$, $p = .006$, $d = .10$, 95% CI = [.03, .17]) trials, but not on the second ($\chi^2 = .02$, $p = .902$, $d = .00$, 95% CI = [−.06, .07]) and fifth ($\chi^2 = 1.47$, $p = .225$, $d = .04$, 95% CI = [−.03, .11]) trials. However, the 95% CI's were overlapping for all trial positions.

[5] We tested whether the length of participants' responses to the power prime (measured as the number of characters in their response) moderated the effect of high versus low-power conditions on sincerity ratings. However, we did not find a reliable Condition × Response Length interaction, $F(1, 2961) = 0.39$, $p = .53$, $r = .01$.

[6] Additional analyses controlling for participants' mood and for task difficulty did not change the direction of the effects, though controlling for mood did weaken the effect ($p = .095$).

[7] The clipboard in the original study was metal. Three of the present sites used a metal clipboard that was very similar to the original, and the rest used a plastic clipboard. The average effect size for the metal clipboard sites was $d = .07$, 95% CI = [−.14, .28], (three sites separately: .09, .02, .12), and the average effect size for the plastic clipboard sites was $d = .02$, 95% CI = [−.07, .11]. The effect was not observed for either clipboard type.

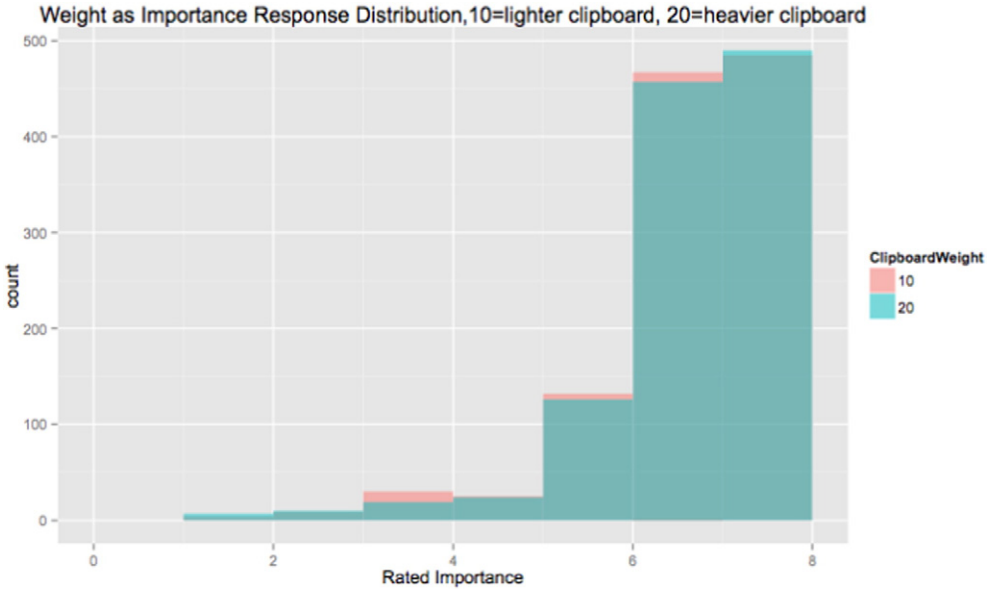

Fig. 2. Histogram plot of importance ratings for Weight Embodiment separated by lighter and heavier clipboard conditions.

description ($M = 66.11$, $SD = 4.34$), $t(78) = 3.85$, $p < .001$, $d = .86$, 95% CI = [.40, 1.33]. This suggests that metaphorically-based conceptualizations of warmth can influence perceptions of the physical environment.

Across the replication studies ($N = 3{,}119$), participants estimated that the ambient room temperature was about the same warmth following the communal description ($M = 71.42$, $SD = 4.97$) as in the agentic description ($M = 71.38$, $SD = 4.79$), $t(3117) = 0.22$, $p = .827$, $d = .01$, 95% CI = [−.06, .08]. The replications did not show evidence of the original effect.[8]

4.8. Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses (Cacioppo, Petty, & Morris, 1983, study 1)

Cacioppo et al. (1983) investigated the impact of argument strength on persuasion, inviting participants who scored in the upper or lower third on the Need for Cognition Scale (Cacioppo & Petty, 1982) to participate. Participants either read a set of strong or weak arguments concerning the institution of comprehensive exams for undergraduates at their university. Afterwards, participants rated the quality of the arguments and how persuaded they were by them. They found that participants found stronger arguments to be more compelling than weaker arguments overall, $F(1, 110) = 160.86$, $p < .001$, $\eta_p^2 = .59$, 95% CI = [.47, .67]. However, participants who were high in need for cognition showed this effect more strongly than those low in need for cognition, $F(1, 110) = 22.45$, $p < .001$, $\eta_p^2 = .17$, 95% CI = [.06, .29]. This study demonstrated that the quality of a persuasive message impacts people differently depending on the extent to which they process the message.

We conducted a similar test using linear regression to predict ratings of argument quality (scored as the average of five follow-up questions about the article) from article condition (strong arguments vs. weak arguments), Need for Cognition (centered), and the Condition × Need for Cognition interaction. Across the in-lab replication studies ($N = 2{,}365$),[9] participants found stronger arguments to be more compelling than weaker arguments, $F(1, 2361) = 79.925$, $p < .001$, $\eta_p^2 = 0.033$, 95% CI = [.020, .048].[10] However, unlike the original study, the interaction term was not reliably different from zero, $F(1, 2361) = 0.129$, $p = .720$, $\eta_p^2 = 5.46$ e-5, 95% CI = [0, .002].[11] Participants' need for cognition did not qualify the effect of argument quality on persuasion.

The reliability of the need for cognition scale used ($\alpha = .67$), was lower than has been observed for the full 34 item scale ($\alpha = .87$, Cacioppo & Petty, 1982). This reduction in reliability would be expected to attenuate the target effect. However, given the statistical power of the sample, it is unlikely that this attenuation would solely eliminate the effect. It could also be that low need for cognition participants is more against comprehensive exams at baseline. However, need for cognition was not reliably related to ratings of argument quality, $F(1, 2361) = 2.386$, $p = .123$, $\eta_p^2 = .001$, 95% CI = [0, .005].

4.9. It feels like yesterday: self-esteem, valence of personal past experiences, and judgments of subjective distance (Ross & Wilson, 2002, study 2)

According to the theory of temporal self-appraisal, time is a psychological variable that can vary by "closeness." Closeness refers to an individual's perception of the temporal distance between the past and the present irrespective of the *actual* temporal distance. For example, a person may have gotten married 15 years ago, but that experience might "feel like" it occurred much more recently. Ross and Wilson (2002) examined how subjective temporal distance varies when recalling negative compared to positive events and whether differences

[8] We constructed a hierarchical multivariate model testing the effect of the manipulation (reading about a communal or agentic individual) with the additional predictors of gender of the individual in the prompt, participant gender, actual room temperature (step 1), and the interaction between target and participant gender (step 2) predicting the participant's temperature estimate of the room. Only the actual room temperature reliably predicted the participants' temperature estimation, $F(1, 1{,}824) = 160.74$, $p < .001$, $r = .28$. All other predictors were not significant ($ps > .41$).

[9] Due to a technological issue, roughly 200 participants from one collection site saw the arguments for this effect from another site. That is, the arguments for institution comprehensive exams were phrased as a proposal at a school not their own. These participants were removed from the analyses.

[10] The main effect of article condition was much smaller in this investigation compared to the original study ($\eta_p^2 = 0.033$ compared to $\eta_p^2 = 0.59$). It could be that the attenuated main effect here diminished the ability to detect the interaction compared to the original. However, across the 20 sites, main effect strength was actually somewhat negatively correlated with interaction effect strength, $r(18) = -.39$, $p = .087$, which speaks against this possibility.

[11] Unlike the original study, we treated Need for Cognition as a continuous variable. Replicating the original analysis using only participant from the upper and lower thirds of the scale reveals similar results, failing to replicate the interaction, $F(1, 1563) = 1.102$, $p = .294$, $\eta_p^2 = 0.0007$, but retaining the main effect of argument condition, $F(1, 1563) = 60.171$, $p < .001$, $\eta_p^2 = 0.023$ with a weaker effect size than when using the whole sample.

in self-esteem may be associated with how distant events subjectively feel, irrespective of how distant they actually are. Overall, participants were expected to feel further from negative events compared to positive events in order to buffer their self-worth against the implications those negative events have for current self-view. Because individuals with high self-esteem are more motivated to preserve their self-worth, the authors hypothesized that individuals with high self-esteem would show this effect more strongly than individuals with low self-esteem.

They randomly assigned students ($N = 357$) to reflect either on a positive or negative academic experience. In the positive condition, participants identified the best grade they received in the previous semester. In the negative condition, participants identified the worst grade they received in the previous semester. Participants then reported how distant the course felt to them and how often they thought about this course since it ended. From a hierarchical regression model (with actual time since the class as step 1, the main effects of self-esteem and condition as step 2, and the interaction of self-esteem and condition as step 3) there was an interaction between self-esteem and condition when predicting ratings of subjective distance, $t(352) = 1.98$, $p = .0485$, $\beta = -.136$, $d = 0.21$, 95% CI = [.001, .418]. Participants who scored high in self-esteem felt more distant from courses in which they received their worst grade, $t(352) = 3.57$, $\beta = -.31$, $p = .0002$. Low self-esteem participants exhibited no significant relation between grade and subjective distance, $t(352) = 0.83$, $\beta = -.07$, $p = 0.204$.

For the aggregate replication test ($N = 3{,}136$), we constructed a hierarchical regression model predicting subjective distance, with actual time since the course (centered) entered in step 1, self-esteem[12] (centered) and condition (best grade or worst grade) entered in step 2, and finally the Self-Esteem × Condition interaction entered in step 3. In an aggregate analysis of all replication studies, we did not detect a reliable Self-Esteem × Condition interaction, $F(1, 3131) = 1.98$, $p = .160$, $\eta_p^2 = .001$, 95% CI = [0, .004]. We did, however, observe the main effect of condition. Participants in the best grade condition felt slightly closer to the recalled class than those in the worst grade condition, $F(1, 3131) = 33.24$, $p < .001$, $\eta_p^2 = .011$, 95% CI = [.005, .019]. Also, self-esteem weakly predicted subjective distance, independent of condition, with higher self-esteem predicting closer subjective distance, $F(1, 3{,}131) = 7.97$, $p = .004$, $\eta_p^2 = .003$, 95% CI = [.0002, .007].

Course grades might have been more relevant to undergraduates compared to MTurk workers. As such, we repeated these analyses for the two groups separately. The results were similar across groups. Neither sample showed the predicted interaction, (for undergraduates: $F[1, 2557] = 1.189$, $p = .276$, $\eta_p^2 = .0005$, 95% CI [0, .004], for MTurk sample: $F[1, 569] = .643$, $p = .423$, $\eta_p^2 = .0011$, 95% CI [0, .013]), and both samples demonstrated the main effect of recall condition (for undergraduates: $F[1, 2557] = 29.809$, $p < .001$, $\eta_p^2 = .012$, 95% CI [.005, .021], for MTurk sample, $F[1, 569] = 4.432$, $p = .036$, $\eta_p^2 = .007$, 95% CI [0, .028]).

4.10. Moral credentials and the expression of prejudice (Monin & Miller, 2001, study 1)

Monin and Miller (2001) tested whether participants were more willing to express prejudicial attitudes when their prior behavior provided evidence that they were non-prejudiced. Two hundred two undergraduates (115 men and 87 women) were approached on a university campus by the experimenter to complete an anonymous survey. This survey first asked whether five statements were right or wrong. These statements expressed sexist views, and were either phrased as describing "most women" or "some women," with the intent of inducing greater agreement with the "some" statements as opposed to the "most" statements. The authors predicted that disagreeing with sexist statements would establish an individual's moral credentials, allowing them to be more prejudiced on subsequent judgments. They did not predict an interaction with gender. Participants in a third condition saw no statements and just completed the remaining measures. Next, participants completed a 3-item filler task before reading a vignette about a manufacturing company that is hiring for a new position. After reading about the position, participants indicated whether or not they believed that the position is better suited for one gender over the other.

Monin and Miller (2001) observed a main effect of moral credentials, $F(2, 194) = 4.4$, $p = .014$, $\eta_p^2 = .043$, 95% CI = [.002, .103], such that participants in the "most" condition ($M = 4.8$) favored a man more than those in the "some" ($M = 4.3$) and base-rate ($M = 4.5$) conditions. There was also a main effect of gender, $F(1, 194) = 9.9$, $p = .002$, and an (unexpected) significant Gender × Credentials interaction, $F(2, 194) = 3.7$, $p = .027$, $\eta_p^2 = .04$, 95% CI = [0, .09]. Only men were influenced by the manipulation. Among men, participants in the "most" condition ($M = 5.1$) differed from both the "some" ($M = 4.4$) and base-rate conditions ($M = 4.6$), $t(54) = 3.3$, $p = .002$, $d = .87$, and $t(86) = 2.7$, $p = .008$, $d = .61$. Men in the "most" condition showed a stronger tendency to endorse men for the job compared to men in the other conditions. The "some" and base-rate conditions did not differ ($t[82] = -1.3$, $p = .197$). There were no significant differences between women in the three conditions ($M = 4.4$, 4.3, and 4.4), all $ts < 1$.

For the replication design, we included only the "some" and "most" conditions. In an aggregate analysis of all replication studies ($N = 3134$), there was a main effect of moral credentials, $F(1, 3130) = 17.01$, $p < .001$, $\eta_p^2 = .005$, 95% CI = [.001, .012]) such that participants in the "most" condition ($M = 4.44$, $SD = 0.94$) favored a man more than participants in the "some" condition ($M = 4.31$, $SD = 0.83$). There was a main effect of gender, $F(1, 3130) = 48.36$, $p < .001$, but not a Gender × Credentials interaction, $F(1, 3130) = 0.0004$, $p = .985$, $\eta_p^2 = .000$.[13] In sum, the replications showed a similar main effect of the predicted credentials manipulation on making sexist judgments, but did not replicate the unexpected moderation by gender.[14]

5. Results by site, task order, and time of semester

For each data collection site, we computed the number of days in which the participant pool was available during the semester. For each participant, their participation date was normalized by dividing the day that they participated by the total number of days available such that participation on the first day of the pool was (1/total days) and participation on the last day of the pool was 1 (see Fig. S1 for distribution of participation). This accounted for the fact that some participant pools were open for longer periods than others (e.g., sites using semesters compared to quarters). This value was tested as a moderator of the association of effect of condition for each of the outcome measures in the study.

5.1. Effects

The primary aim of the pre-registered design and analysis plan was to evaluate variation in effects across the academic semester. In the first stage of analysis, we examined the aggregate effect sizes without testing whether those effects varied across the semester. Those results, reported above in the introduction to each effect, suggested that some of the primary replication effects had effect sizes near 0. It was possible that this aggregate result would reveal a positive effect at some points in time and a negative effect at other points in time. However, it was also

[12] We used a single item self-esteem measure instead of the Rosenberg Self-Esteem Scale (1965), which was used in the original study.

[13] The effect size for this interaction was estimated as 1.28 e-6. The weighted upper bound of the 95% CI was too small to compute with the statistical software.

[14] A secondary dependent measure assessed whether participants agreed that "women are just as able as men to do any kind of job?" (−3 "strongly disagree" to 3 "strongly agree). Using the same analysis plan as for the primary replication, we did not find a main effect of credential condition or a Condition × Gender interaction. Unsurprisingly, there was a main effect of gender, with women ($M = 1.89$, $SD = 1.30$) agreeing with the statement more than men ($M = 1.03$, $SD = 1.62$,), $F(1, 3{,}127) = 255.30$, $p < .001$, $r = .27$.

possible that this indicated a uniformly null result. If the latter, then we would have no opportunity to learn about variation across the academic semester from those effects. As a consequence, prior to conducting tests of variation across time, we decided to add three theoretically relevant main effects for studies in which the key test was an interaction effect that did not occur (Elaboration Likelihood, Self-Esteem and Subjective Distance, Credentials and Prejudice).

5.1.1. Variation by site

For each effect, we computed an aggregate effect size estimate with 99% confidence intervals. Fig. 1a and b represents the effect size estimates for each of the data collection sites for each effect. We also computed the variability in effect estimates following standard statistics for meta-analyses—Q and I^2—to determine if the amount of variability across samples exceeds that expected by random error. With identical study procedures, variability exceeding expectations of sampling error is likely attributable to variation in the effect due to sample or setting. These analyses are presented in Table 4. Overall, two effects, Self-Esteem and Subjective Distance and Credentials and Prejudice, showed signs of inter-site variation. For both effects, the interactions (I^2 = 39.25%, $p = .026$; $I^2 = 28.17\%$, $p = .068$, respectively) and main effects ($I^2 = 35.81\%$, $p = .063$; $I^2 = 40.31\%$, $p = .023$, respectively) showed small to moderate variation, according to meta-analytic standards (Higgins, Thompson, Deeks, & Altman, 2003). All other effects showed little inter-site variation ($Q < 22.40$, $p > .288$).

5.1.2. Variation by task order

Across the session, effects may weaken if participants get fatigued or if prior measures interfere with subsequent measures. To investigate this possibility, we conducted moderator analyses on each of the 10 + 3 effects, testing for linear and quadratic order effects (see Table 5 for summary and Table S6 for other tests of order effects). Overall, we observed very little variation by task order (average $\eta_p^2 = .0002$ for effects with non-binomial outcomes, average $d = .04$ for effects with binomial outcomes). In addition, we analyzed the data from each effect when it was presented first in the task sequence. Comparing these results to the aggregate results revealed little variation. Metaphoric Restructuring was slightly weaker when presented first ($\Delta d = -.18$) and Availability was slightly stronger ($\Delta d = .17$). The Elaboration Likelihood main effect was also slightly stronger ($\Delta\eta_p^2 = .04$) when presented first. All other effects showed similar strength.

5.1.3. Variation by time of semester

Our primary interest was in the variation of effects across the academic semester. For each of the 10 replicated effects (and 3 additional main effects) we first constructed an unconditional model, predicting the outcome variable from a fixed intercept and a random intercept of site. This was to determine the amount of variation in outcome variables between sites before examining time of semester variation in effect detection. For all but two of the models, site accounted for 1.1% of the variance or less in the dependent variable. There were non-trivial site effects for the persistence measure (5.0%; Persistence and Conscientiousness) and for temperature (22%; Warmth Perceptions). Students at some sites were more persistent with the anagrams than at other sites, and some lab rooms were perceived as warmer than others. Otherwise, there was little variation in the dependent variables by site.

Then, we constructed a mixed effects model, with the Time of Semester × Replication Independent Variable(s) as a fixed effect. We included a random intercept of site and random slope of the fixed effect by site. For many models, this random slope overparameterized the model, and was thus simplified or dropped. The final model for each effect was compared to a model without Time of Semester as a fixed interaction to test whether adding Time of Semester provided a better fit for the data.[15] We performed these analyses on participant pool participants first, and we planned to use the MTurk sample as a comparison when time of semester variation was observed. See Table 6 for a summary of variation by semester analyses.

There was little evidence for variation by time of semester for most effects. Of the 13 tested effects, model fit comparisons provided very weak evidence for three effects with slight differences by time of semester, Stroop ($p = .069$), Warmth Perceptions ($p = .049$), and Metaphoric Restructuring ($p = .034$). Variation in Warmth Perceptions was due to slight evidence of a main effect of temperature ratings declining over the course of the semester, presumably because of the onset of Fall ($\eta_p^2 = .001$, 95% CI = [0, .005]), and variation in Stroop was evidence of a slightly *stronger* Stroop effect later in the semester ($\eta_p^2 = .002$, 95% CI = [0, .007]). For variation of Metaphoric Restructuring, we conducted a logistic regression predicting prime consistent responding from Time of Semester but found no significant effect $\chi^2(1, N = 1332) = 0.010$, $p = .920$. However, this effect did show variation near the end of the semester. We observed a slightly *stronger* effect in the last 20% of the semester ($d = .36$) compared to the first 80% ($d = .24$). MTurk participants showed directionally similar but unreliable patterns, including a very small increase in the Stroop effect over time, $F(1, 618) = 2.25$, $p = .134$, $\eta_p^2 = .004$, 95% CI = [0, .019], and tiny declining temperature estimates over time $F(1, 571) = .88$, $p = .350$, $\eta_p^2 = .002$, 95% CI = [0, .014]. We did not administer the Metaphoric Restructuring task to the MTurk participants.

With so little evidence for a time of semester effect, we conducted a follow-up exploratory analysis comparing data from the first 80% of the semester to the last 20% of the semester. This was to focus the test on the intuition that the inattentive or unmotivated participants are those that complete studies at the very end of the semester. Results are summarized in supplementary Table S7. Again, we found little evidence of variation in effect magnitudes, observing the largest difference for Metaphoric Restructuring. The large number of comparisons suggests caution in interpreting this effect, however.

Overall, the data revealed little evidence for variation in effect magnitudes by time of semester. Even when just considering effects that replicated in aggregate, only two of six effects showed hints of time of semester variation (Stroop, Metaphoric Restructuring). In both cases, the effects were actually larger toward the end of the semester compared to the rest of the semester.

Table 4
Heterogeneity in effect sizes by data collection site.

	Heterogeneity tests			
Effect	Q	df	p-value	I^2
Stroop task	15.1883	20	0.7655	4.05%
Metaphoric restructuring	21.9213	19	0.2882	18.23%
Availability heuristic	19.9805	20	0.4591	1.21%
Persistence and conscientiousness	22.4037	20	0.319	1.40%
Power and perspective	19.7975	20	0.4707	0.01%
Weight embodiment	12.2518	19	0.8746	<0.005%
Warmth perceptions	16.9429	20	0.6567	<0.005%
Elaboration likelihood	11.3995	19	0.9097	<0.005%
Self-esteem and subjective distance	34.0701	20	0.0257	39.25%
Credentials and prejudice	30.1262	20	0.0678	28.17%
*Elaboration likelihood – Main effect	12.6038	19	0.8582	<0.005%
*S-E and subjective distance – Main effect	30.4253	20	0.0632	35.81%
*Credentials and prejudice – Main effect	33.0233	20	0.0233	40.31%

Note. Effects were ordered from the largest to the smallest observed effect size (see Table 3). Heterogeneity tests conducted with R-package metafor. REML was used for estimation for all tests.

[15] The mixed effects model for Weight Embodiment violated the assumption of normally distributed residual variance. To correct this, we inverse reflection transformed the response variable using the formula: 1/(8 - response).

Table 5
Order effects by task order.

Effect	F (linear)	df (interaction)	df (residuals)	p-value	η_p^2	95% CI	F (quadratic)	df (interaction)	df (residuals)	p-value	η_p^2	95% CI
Stroop task	1.60	1	3278	0.21	0.00049	0, 0.003	2.38	1	3278	0.12	0.00073	0, 0.004
Persistence and conscientiousness	0.02	1	3189	0.88	0.00001	0, 0.001	0.06	1	3189	0.81	0.00002	0, 0.001
Power and perspective	1.17	1	2965	0.28	0.00039	0, 0.003	0.54	1	2965	0.46	0.00018	0, 0.002
Weight embodiment	0.02	1	2066	0.88	0.00001	0, 0.002	0.00	1	2066	0.95	0.00000	0, 0.001
Warmth perceptions	0.05	1	3115	0.82	0.00002	0, 0.001	0.12	1	3115	0.72	0.00004	0, 0.002
Elaboration likelihood	0.35	1	2357	0.56	0.00015	0, 0.003	0.04	1	2357	0.84	0.00002	0, 0.002
Self-esteem and subjective distance	0.42	1	3127	0.52	0.00013	0, 0.002	0.60	1	3127	0.44	0.00019	0, 0.002
Credentials and prejudice	0.02	1	3126	0.90	0.00000	0, 0.001	0.08	1	3126	0.78	0.00003	0, 0.001
*Elaboration likelihood – Main effect	2.04	1	2357	0.15	0.00087	0, 0.005	2.08	1	2357	0.15	0.00088	0, 0.005
*S-E and subjective distance – Main effect	0.09	1	3127	0.76	0.00003	0, 0.001	0.43	1	3127	0.51	0.00014	0, 0.002
*Credentials and prejudice – Main effect	0.33	1	3126	0.57	0.00010	0, 0.002	0.10	1	3126	0.76	0.00003	0, 0.001
Averages	0.56			0.59	0.00020		0.58			0.59	0.00021	

Binomial outcomes	Likelihood Chi-square (linear)	p-value	d	95% CI	Likelihood Chi-square (quadratic)	p-value	d	95% CI
Metaphoric restructuring	0.03	0.87	0.01	−0.10, 0.12	0.17	0.68	0.02	−0.09, 0.14
Availability heuristic	3.11	0.08	0.06	−0.01, 0.13	2.17	0.14	0.05	−0.02, 0.12
Averages	1.57	0.48	0.04		1.17	0.41	0.04	

5.2. Data quality indicators

Participants reported fairly high levels of effort ($M = 3.71$, $SD = .78$; Scale 1 = *no effort* to 5 = *tried my hardest*) and attention ($M = 3.92$, $SD = .74$; Scale 1 = *none* to 5 = *I gave my undivided attention*), and 37.3% failed the instructional attention check, similar to prior demonstrations with this challenging check (Oppenheimer, Meyvis, & Davidenko, 2009). Participants demonstrated some awareness of their attention levels. Participants who passed the attention check reported higher attention ($M = 4.03$, $SD = .70$) than those who failed the check ($M = 3.78$, $SD = .77$), $t(2606) = 8.20$, $p < .001$, $d = .33$, 95% CI = [.25, .41].

Table 6
Moderation of effect sizes by time of semester.

Effect	Variation in outcome by site (R2)	Overall model fit time of semester	p-value	Time of semester	p-value	Partial eta-sq	95% CI
Stroop task	0.6%	$\chi2(1, N = 2660) = 3.31$	0.069	$F(1, 2658) = 5.01$	0.025	0.002	0, 0.007
Metaphoric restructuring	0.01%	$\chi2(1, N = 1332) = 4.48$	0.034	$\chi2(1, N = 1332) = .010$	0.92		
Persistence and conscientiousness	5.0%	$\chi2(2, N = 2624) = 4.63$	0.099				
Availability heuristic	0.01%	$\chi2(4, N = 2497) = 1.45$	0.228				
Power and perspective	0.9%	$\chi2(2, N = 2385) = .70$	0.699				
Weight embodiment	1.7%	$\chi2(2, N = 2279) = 3.97$	0.138				
Warmth perceptions	22.0%	$\chi2(2, N = 2544) = 6.04$	0.049	$F(1, 1842) = 3.83$	0.051	0.002	0, 0.008
Elaboration likelihood	1.1%	$\chi2(4, N = 2365) = 2.02$	0.732				
Self-esteem and subjective distance	0.9%	$\chi2(4, N = 2562) = .54$	0.969				
Credentials and prejudice	0.4%	$\chi2(4, N = 2571) = 4.90$	0.298				
*Elaboration likelihood – Main effect	1.1%	$\chi2(2, N = 2429) = .22$	0.896				
*S-E and subjective distance – Main effect	0.9%	$\chi2(2, N = 2562) = .32$	0.851				
*Credentials and prejudice – Main effect	0.4%	$\chi2(2, N = 2642) = 2.15$	0.341				
Data quality indicators							
Attention check	4%	$\chi2(1, N = 2621) = 6.75$	0.009	$r(2621) = -.08$	<.001		−.12, −.04
Reported effort	2.5%	$\chi2(1, N = 2628) = 17.46$	<.001	$r(2626) = -.11$	<.001		−.14, −.07
Reported attention	1.6%	$\chi2(1, N = 2630) = 11.60$	<.001	$r(2628) = -.08$	<.001		−.12, −.04
Demographics							
Age	2.6%	$\chi2(1, N = 2592) = 0.05$	0.821				
Sex	3.7%	$\chi2(1, N = 2598) = 17.57$	<.001	$r(2598) = 0.12$	<.001		.08, .16
Race/ethnicity	1.7%	$\chi2(1, N = 2607) = 2.38$	0.123				
Year in college	14.2%	$\chi2(1, N = 2570) = 0.89$	0.346				
Individual differences							
Conscientiousness	4.2%	$\chi2(1, N = 2628) = 32.11$	<.001	$r(2626) = -.14$	<.001		−.18, −.10
Agreeableness	<0.01%	$\chi2(1, N = 2629) = 0.005$	0.945				
Extraversion	2.2%	$\chi2(1, N = 2625) = 2.40$	0.121				
Neuroticism	1.1%	$\chi2(1, N = 2630) = 1.31$	0.252				
Openness to experience	1.8%	$\chi2(1, N = 2631) = 0.01$	0.923				
Intrinsic motivation	1.2%	$\chi2(1, N = 2608) = 0.14$	0.71				
Stress	1.9%	$\chi2(1, N = 2623) = 10.08$	0.001	$r(2621) = .08$	<.001		.04, .12
Mood	1.2%	$\chi2(1, N = 2636) = 8.03$	0.005	$r(2634) = -.07$	0.001		−.10, −.03
Self-esteem	0.4%	$\chi2(1, N = 2625) = 0.56$	0.456				
Need for cognition	1.1%	$\chi2(1, N = 2601) < 0.00005$	0.998				

Note. Variation by site indicates the amount of variation in the dependent variable attributable to location of data collection. Follow-up tests of time of semester predicting variation in the effect conducted for only those effects in which the overall model improved ($p < .07$) by adding time of semester as a factor. Two of the outcomes, sex and attention check, had binary outcomes. Changes over time in those variables are quantified by odds ratio of a given outcome on the last day of the semester compared to the first day of the semester (odds ratio estimates taken from the mixed model).

Only one effect, Availability Heuristic, was reliably moderated by performance on the attention check, with those who failed the check actually showing a stronger effect ($p = .032$, $d = .08$; see Table S8 for a full summary of results). The attention check did not moderate any of the time of semester effects observed either ($ps > .512$).

To analyze time of semester variation in these data quality indicators, we constructed mixed effects models predicting the data quality indicators with Time of Semester as a fixed effect and a random intercept of Site (see Table 6). We compared these models to models without Time of Semester as a fixed effect. Unconditional models revealed that 2.5% of variance in Reported Effort, 1.6% of the variance in Reported Attention, and 4% of the variance in the Attention Check was explained by inter-site variation. This suggests more variation in effort and attention across sites than variation in responses on most of the dependent variables.

Model comparison revealed that the addition of Time of Semester reliably improved model fit for Reported Effort ($p < .001$), Reported Attention ($p < .001$), and the Attention Check ($p = .009$). As the semester progressed, reported effort declined, $r(2626) = -.11$, $p < .001$, 95% CI = [−.14, −.07], as did reported attention, $r(2628) = -.08$, $p < .001$, 95% CI = [−.12, −.04], and participants were more likely to fail the attention check, $r(2621) = -.08$, $p < .001$, 95% CI = [−.12, −.04]. All of these effects were small.

5.3. Demographics

To observe demographic trends over the semester, we constructed mixed effects models predicting participant sex, age, ethnicity, and year in school from Time of Semester as a fixed effect and a random intercept of Site (see Table 6 for a summary). Time of Semester only reliably improved the model for participant sex, $\chi^2(1, N = 2598) = 17.57$, $p < .001$. Participants were more likely to be male as the semester progressed, $r(2598) = .12$, $p < .001$, 95% CI = [.08, .16].

5.4. Individual differences

To evaluate variation across the semester, we constructed linear mixed effects models testing each of the 10 individual difference variables (see Table 6). We compared a model with Time of Semester as a fixed effect and a random intercept of Site with a model lacking the Time of Semester fixed effect. These model comparisons revealed that Time of Semester reliably improved models for conscientiousness ($p < .001$), mood ($p = .005$), and stress ($p = .001$). Follow-up analyses showed that as the semester progressed, participants were less conscientious, $r(2626) = -.14$, $p < .001$, 95% CI = [−.18, −.10], reported worse mood, $r(2634) = -.07$, $p = .001$, 95% CI = [−.10, −.03], and reported being more stressed, $r(2621) = .08$, $p < .001$, 95% CI = [.04, .12]. All of these effects were small, and none of the other individual differences were reliably moderated by time of semester.

Finally, in exploratory analyses, we investigated whether the data quality indicators or individual differences that varied over the semester moderated the effects that varied over the semester (Stroop, Metaphoric Restructuring). However, none of these data quality indicators or individual differences moderated Stroop or Metaphoric Restructuring (all p's > .253, see supplementary materials for details).

6. Discussion

This crowdsourced project evaluated whether variation in effect magnitudes can be partially attributed to the time of semester of data collection. The answer from the 10 + 3 investigated effects is largely no. Detected effects had similar effect sizes regardless of when data collection occurred and effects that were not detectable during some part of the semester were not detectable at any point during the semester.

Consistent with literature showing that Stroop effects are sensitive to the availability of cognitive resources to overcome response competition (Kane & Engle, 2003), the Stroop effect was slightly stronger toward the end of the semester (last 20% $d = .92$) compared to the beginning (first 80% $d = .89$), but even that effect was very small. Also, there was a hint of stronger effects for Metaphoric Restructuring at the end of the semester compared to earlier. All told, effects showed little to no moderation by time of semester, site of data collection, and order in which the tasks were administered.

Qualifying the generality of the conclusion, of the ten original effects we examined, only three replicated the original result, regardless of the time of semester. After observing this, but prior to testing time of semester effects, we added three successful main effect replications. These provided no additional evidence for time of semester effects. Examining only the reliable effects, two of the six showed any time of semester variation, and those two effects became very slightly stronger, not weaker, in the latter parts of the semester.

The conclusions would be more definitive had a greater proportion of the effects shown a reliable result. Moreover, the selection of effects was by no means a random selection or representative sample of all possible effects. As such, the present results provide a provocative, but constrained conclusion. With a very high-powered design, time of semester was largely irrelevant for estimating the magnitude of experimental and correlational effects. The extent to which the present results will generalize across replicable experimental and correlational effects is unknown.

6.1. What does change across the academic semester

If effects do not change across the semester, what does? The present study replicated and extended prior observations (Nicholls et al., 2015; Witt et al., 2011). As the semester progressed, participants reported slightly less effort and attention, were slightly more likely to fail an attention check, were slightly less conscientious, had slightly worse mood, had slightly higher stress, and had slightly higher representation of men compared to women. These effects are regularly hypothesized and easily recognized by frequent users of participant pools even though time of semester accounts for only about 1% of the variance in each. As such, participant characteristics did shift slightly across the semester, but these shifts had little impact on the detectability of the tested correlations and experimental results. In fact, these indicators suggest slightly weakening data quality later in the semester, but the two effects that did change actually showed stronger effects toward the end.

6.2. Moderation of effects

A common explanation for the challenges of replicating results across samples and settings is that there are many seen and unseen moderators that qualify the detectability of effects (Cesario, 2014). As such, when differences are observed across study administrations, it is easy to default to the assumption that it must be due to features differing between the samples and settings. Besides time of semester, we tested whether the site of data collection, and the order of administration during the study session moderated the effects. Whether the task was administered first, in the middle, or last had minimal impact on the investigated effects. This is consistent with the first “Many Labs” study (Klein et al., 2014). This suggests against the possibility that there is something about the procedure of combining studies into a single session that disrupts detectability of effects.

We did observe some evidence of variation by sample or setting for main effects and interactions of two of the ten studies, Self-Esteem and Subjective Distance and Credentials and Prejudice. These are demonstrations of a truism in social psychology – that effects vary by sample and setting. If anything, it is notable that sample and setting variation was not more prevalent. Investigating variation by sample and setting is the focus of the second “Many Labs” study with many samples and societies included in the study (Klein et al., 2015).

Another potential moderator of well-known effects is participant knowledge. For example, Elaboration Likelihood and Availability Heuristic are often taught in introductory psychology classes. If students learned about these effects in their courses, it is possible that this would reduce observed effects (see, however, Lambdin and Shaffer (2009)). However, if that were the case, we would expect to observe time of semester variation on classic effects, as students would presumably learn about these effects sometime during the academic term, making them less detectable near the end of the term. In addition, students would likely vary in their knowledge of these effects from site to site, as different lessons would be taught at different universities. Given the lack of variation from these two sources this seems unlikely to have occurred.

6.3. Insights about the selected effects

We were surprised that several effects showed null effects in our large sample. The present study's very large sample size and lack of moderating effects by site, order, and time of semester does provide precision and some definitiveness about these paradigms under these conditions. However, *under these conditions*, is a critical qualifying phrase. The present results do not definitively suggest that the observed nulls are always null, nor do they definitively suggest that the original positive results are false positives. What can be concluded is that those effects are not distinguishable from zero with the samples, settings, materials, and procedure employed here.

Even among effects that did replicate in aggregate, we observed smaller effect sizes compared to the original demonstrations. Although past replication projects have observed similar declines in effect strength (Open Science Collaboration, 2015), there are several possible explanations for the declines observed in this study. For instance, in some cases, the original materials or methods were altered to accommodate the constraints of this investigation.[16] Based on a priori theorizing and review these alterations were not expected to alter the results of the replications substantially.[17] Even so, such changes might have had unexpected influences. For example, the original weight importance study was conducted with a Dutch sample. Our samples were from North America, and we did not anticipate this change to qualify the effect based on the current theoretical understanding. Across our 20 sites, we observed mean scores in both conditions of the weight importance study that were within a point of the ceiling reducing the power to detect an effect. The original study had lower means particularly in the light clipboard condition making it an outlier by comparison. It is possible that the change in samples is responsible for the shift in means. Another possibility is that the original study's lower means, particularly in one condition, were an unusual chance occurrence. Parsing between these possibilities requires conducting a replication that includes the original (Dutch) population.

Furthermore, many of the theories invoked by the selected effects have been demonstrated using various methods. For our purposes, we had to select a single instantiation. Thus, our results can only speak to those instantiations, not necessarily the broader theory. A better theoretical understanding of each effect, and the theories they are derived from, will be achieved when the conditions for influencing the effect magnitude are articulated and demonstrated empirically. The present evidence and further explorations of the dataset may provide useful hypothesizing for how to begin that search.

A notable procedural difference between this and the original studies is that the effects were investigated in a single experimental protocol. It is reasonable to hypothesize that this procedural difference produced smaller effects in the replications particularly if they occur later in the ~30 min protocol. The present evidence suggests that this did not occur, particularly because the order of tasks did not moderate the observed effects, including considering only the first task completed. Moreover, the first Many Labs project (Klein et al., 2014) had a similar procedure and reliably detected 10 (and the 11th weakly) of 13 effects, with some effects producing larger effect size estimates compared to the original. An untested possibility is that the procedure weakened effects of even the first task completed because participants anticipated doing many tasks. However, because there were no order effects in this study, such an influence would need to be equal to the disruptive impact of having just experienced the other tasks. While we do not find this idea to be particularly plausible, it would be straightforward to test in new research.

Three of the investigated interaction effects did not replicate: Elaboration Likelihood, Self-Esteem and Subjective Distance, and Credentials and Prejudice. In all three cases, a theoretically relevant main effect was observed. For Credentials and Prejudice, we observed the effect of credentials on prejudice, but that effect was not qualified by gender (Monin & Miller, 2001). Our perception is that the interaction effect is much less theoretically essential than the main effect of credentials. In fact, Monin and Miller (2001) did not anticipate an interaction; it emerged unexpectedly in their first study and did not persist in their second study, which used a different manipulation of credentials. As such, the present replication can be seen as affirming their original theoretical expectations, and disconfirming the unexpected moderation by gender. Likewise, for Self-Esteem and Subjective Distance (Ross & Wilson, 2002), we did observe that positive past events felt closer than negative past events, but we did not replicate the moderation of this effect by self-esteem. In our view, the main effect is most vital theoretically. Finally, in Elaboration Likelihood (Cacioppo et al., 1983), we observed that stronger arguments were more persuasive than weaker arguments, but this was not qualified by need for cognition. The failure to replicate this interaction is one of the most surprising results from this study. The Elaboration Likelihood Model is among the most developed and empirically investigated theories in psychology (Petty & Briñol, 2012). Our result would seem to be an outlier in the literature, albeit a highly precise one. In light of this, it is important to note that we only tested one instantiation relevant to this theory. Post hoc, we examined possible moderators to account for the difference, but did not find support for any of them. We do not have an explanation for why no effect was observed under these circumstances.

6.4. Additional analysis opportunities

The amassed dataset is rich for exploring the individual effects, individual difference variables, interactions between the two, and alternate ways to analyze the aggregate data. Our analysis plan for the main article focused on time of semester variability and not, for example, exploring moderating influences in depth. However, the data set and all materials are available publicly to encourage further investigations by others (visit https://osf.io/ct89g/).

7. Conclusion

Conventional wisdom among behavioral scientists suggests that the time of semester for data collection from participant pools is an important factor for obtaining effects. Our powerful design across 20 participant pools found more evidence against this conclusion than for it. We did find evidence that the characteristics of the sample changes across the semester, but those changes did not alter detection of the selected effects.

Should researchers now discount conventional wisdom? Therein is the incompleteness of any single investigation. The present results are the only known large-scale investigation of the influence of time of semester on a variety of effects. As such, the present results should give pause to speculative invocations of time of semester as an explanatory

[16] Detailed explanations of all known alterations are located in Supplementary Information: Methods for Selected Effects and a summary of alterations can be found in Table S4.

[17] With the exception of the Stroop effect. We used a simplified version with fewer colors and trials than most research applications, so a smaller (still large) effect size was expected.

factor. At the same time, conventional wisdom often has a basis in experience. It is possible that there are some conditions under which the time of semester impacts observed effects. However, it is unknown whether that impact is ever big enough to be meaningful.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jesp.2015.10.012.

References

Amabile, T. M., Hill, K. G., Hennessey, B. A., & Tighe, E. M. (1994). The Work Preference Inventory: assessing intrinsic and extrinsic motivational orientations. *Journal of Personality and Social Psychology*, *66*(5), 950.

Aspinwall, L. G., & Richter, L. (1999). Optimism and self-mastery predict more rapid disengagement from unsolvable tasks in the presence of alternatives. *Motivation and Emotion*, *23*, 221–245.

Aviv, A. L., Zelenski, J. M., Rallo, L., & Larsen, R. J. (2002). Who comes when: Personality differences in early and later participation in a university subject pool. *Personality and Individual Differences*, *33*, 487–496.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, *75*, 1–28.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, *45*, 805–818.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*(1), 40–48.

Cloninger, C. R. (1987). A systematic method for clinical description and classification of personality variants: A proposal. *Archives of General Psychiatry*, *44*, 573–588.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*, 385–396.

De Fruyt, F., Van De Wiele, L., & Van Heeringen, C. (2000). Cloninger's psychobiological model of temperament and character and the five-factor model of personality. *Personality and Individual Differences*, *29*(3), 441–452.

Ehrhart, M. G., Ehrhart, K. H., Roesch, S. C., Chung-Herrera, B. G., Nadler, K., & Bradshaw, K. (2009). Testing the latent factor structure and construct validity of the Ten-Item Personality Inventory. *Personality and Individual Differences*, *47*, 900–905.

Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychological Science*, *17*, 1068–1074.

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test–retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*, *52*, 20–28.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology*, *2*, 141–165.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: univariate and multivariate applications*. New York, NY: Routledge.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ [British Medical Journal]*, *327*(7414), 557–560.

Hom, H. L. (1987). A methodological note: Time of participation effects on intrinsic motivation. *Personality and Social Psychology Bulletin*, *13*, 210–215.

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Science*, *20*, 1169–1174.

Kane, M. J., & Engle, R.W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Alper, S., Aveyard, M., Axt, J. R., ... Nosek, B. A. (2015). *Many labs 2: investigating variation in replicability across sample and setting*. Manuscript in preparation.

Lai, V. T., & Boroditsky, L. (2013). The immediate and chronic influence of spatio-temporal metaphors on the mental representations of time in English, Mandarin, and Mandarin–English speakers. *Frontiers in Psychology*, *4*, 142.

Lambdin, C., & Shaffer, V. A. (2009). Are within-subjects designs transparent? *Judgment and Decision Making*, *4*(7), 554–556.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*, 33–43.

Nicholls, M. E., Loveless, K. M., Thomas, N. A., Loetscher, T., & Churches, O. (2015). Some participants may be better than others: Sustained attention and motivation are higher early in semester. *The Quarterly Journal of Experimental Psychology*, *68*, 10–18.

Nosek, B. A., & Sriram, N. (2007). Faulty assumptions: A comment on Blanton, Jaccard, Gonzales, and Christie (2006). *Journal of Experimental Social Psychology*, *43*, 393–398.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251)http://dx.doi.org/10.1126/science.aac4716.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Petty, R. E., & Briñol, P. (2012). The elaboration likelihood model. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 224–245). London: Sage.

Petty, R. E., Cacioppo, J. T., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). Oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*, 747–758.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem scale. *Personality and Social Psychology Bulletin*, *27*, 151–161.

Rojas, S. L., & Widiger, T. A. (2014). Convergent and discriminant validity of the Five Factor Form. *Assessment*, *21*, 143–157.

Ross, M., & Wilson, A. E. (2002). It feels like yesterday: Self-esteem, valence of personal past experiences, and judgments of subjective distance. *Journal of Personality and Social Psychology*, *82*, 792–803.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*, 513–523.

Sommer, K. L., & Baumeister, R. F. (2002). Self-evaluation, persistence, and performance following implicit rejection: The role of trait self-esteem. *Personality and Social Psychology Bulletin*, *28*, 926–938.

Sriram, N., Greenwald, A. G., & Nosek, B. A. (2010). Correlational biases in mean response latency differences. *Statistical Methodology*, *7*(3), 277–291.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Szymkow, A., Chandler, J., Ijzerman, H., Parzuchowski, M., & Wojciszke, B. (2013). Warmer hearts, warmer rooms. *Social Psychology*, *44*, 167–176.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Verplanken, B. (1991). Persuasive communication of risk information: A test of cue versus message processing effects in a field experiment. *Personality and Social Psychology Bulletin*, *17*(2), 188–193.

Verplanken, B., Hazenberg, P. T., & Palenewen, G. R. (1992). Need for cognition and external information search effort. *Journal of Research in Personality*, *26*(2), 128–136.

Verhaeghen, P., & De Meersman, L. (1998). Aging and the Stroop effect: A meta-analysis. *Psychology and Aging*, *13*(1), 120.

Wang, A. Y., & Jentsch, F. G. (1998). Point-of-time effects across the semester: Is there a sampling bias? *The Journal of Psychology*, *132*, 211–219.

Witt, E. A., Donnellan, M. B., & Orlando, M. J. (2011). Timing and selection effects within a psychology subject pool: Personality and sex matter. *Personality and Individual Differences*, *50*, 355–359.