Supplementary Information: Methods for Selected Effects

Here, we present the materials, procedure, and analysis plan for each of the ten selected effects. These descriptions are taken from the preregistered protocol available at OSF (https://osf.io/5ykuj/). Many of the original studies were conducted with paper and pencil, but most of the replications were conducted via computer. Any other known differences from the original study besides these are noted in the method description.

1. Stroop task (Stroop, 1935)

Materials and procedure. Participants will complete a variation of the Stroop task in which they are presented with the words "red", "blue", and "green" written in either red, blue, or green. Within participants, in one third of trials the word will match the color of the text and in two thirds of the trials the word will not match the color of the text. Participants will read a description of the task and complete a 12-trial practice block before the critical trials. The description of the task is as follows:

"On the following page, you will complete a color recognition task. Several words will be presented. For each word, indicate the color of the font by pressing the appropriate key.

Specifically, if the word is presented in red, press the 1 key. If the word is presented in blue, press the 2 key. If the word is presented in green, press the 3 key.

So for instance, if you saw the word red, you would press the 1 key.

If you saw the word blue, you would also press the 1 key because although the meaning of the word is a different color, the color of the font is still red.

Complete this task as quickly as possible while minimizing errors.

Once you have understood these instructions, place your fingers on the 1, 2, and 3 keys and press 'continue' to begin the task. You will first do a practice block before continuing to the test block."

When participants begin, items appear one at a time in the middle of the screen and labels at the top of the screen remind participants of which key is associated with which color response. If participants make a categorization error, a red X will appear on the screen to note the error, and the task will advance to the next trial. Upon completion of the practice block, participants will be presented with the description of the task again and then complete a 63-trial test block so that each color word appears 21 items, 7 times in each of the three font colors.

Analysis plan. 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. The analysis strategy uses the *D* scoring algorithm for analysis of such data (Greenwald, Nosek, & Banaji, 2003; Nosek & Sriram, 2007). First, all trials with latencies above 10,000ms will be removed. Then, we will calculate an average response time for all *correct* responses separately for congruent and incongruent trials. Response latencies for trials with errors will be replaced by the mean of correct responses in that condition plus 600 milliseconds. Then, the means for congruent and incongruent trials will be recomputed

with all trials. *D* is the difference between these two means divided by the standard deviation of all trials regardless of condition. Positive scores will indicate slower response times on average for incongruent compared to congruent trials.

As a secondary analysis, we will investigate Stroop interference as a function of errors by calculating the difference between the proportion of error in incongruent versus congruent conditions.

Known differences from original. The procedural details for the Stroop vary widely across the many hundreds of research applications. Fortunately, the effect is extremely robust across variations. For example, the original demonstration of this effect (Stroop, 1935) used more color words than the current study, obtaining an effect of $d = 3.07$. More recently, Inzlicht and Gutsell (2007) used a version with two color words and 864 trials ($d = 3.14$). A meta-analysis comparing Stroop performance of young adults to that of older adults found an aggregate effect size of $d = 2.04$ for younger adults (the group most similar to our sample, Verhaeghen & De Meersman, 1998). The current design was chosen in order to make the task simple enough for participants to learn and complete in a short amount of time while still retaining enough complexity to obtain a robust effect. However, given the fewer number of color stimuli and fewer number of trials, we expected a smaller effect size and thus did not include the original as a benchmark for comparison.

2. Metaphoric structuring: Understanding time through spatial metaphors (Boroditsky, 2000, Study 1)

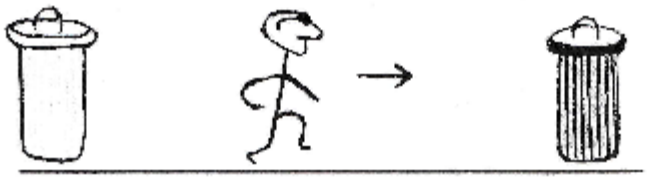
Materials and procedure. Based on the original author's recommendation, this task will be completed on paper-and-pencil in the face-to-face portion of the study to ensure comparability to the original procedure.

Participants will be randomly assigned to the ego-moving, object-moving, or control conditions. For the first two conditions, participants will be asked to decide whether each statement accompanying four images is true or false. The ego-moving condition will include the first four images below, the object-moving condition will include the second four images below. Participants in the control condition will not see or rate any images.

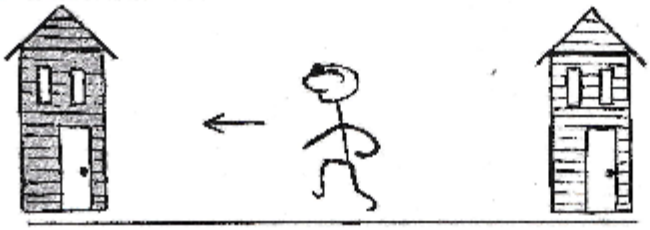
Subject ID #:_____

Please circle TRUE or FALSE for questions 1-4 below.

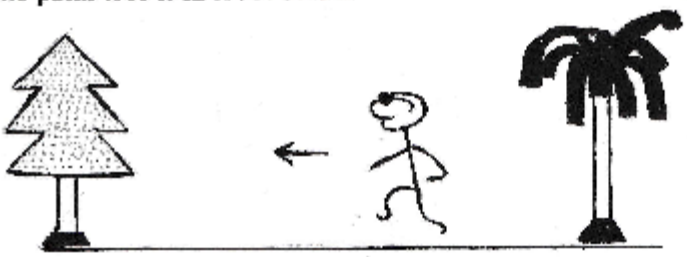
1. The light can is in front of me. **TRUE / FALSE**



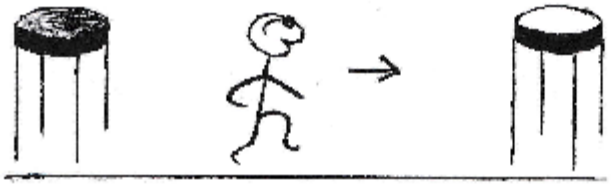
2. The dark house is in front of me. **TRUE / FALSE**



3. The palm tree is in front of me. **TRUE / FALSE**



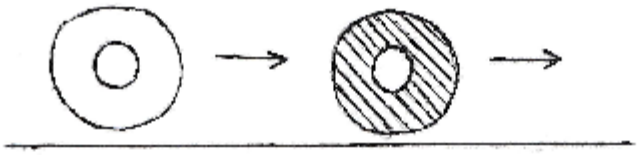
4. The white stool is in front of me. **TRUE / FALSE**



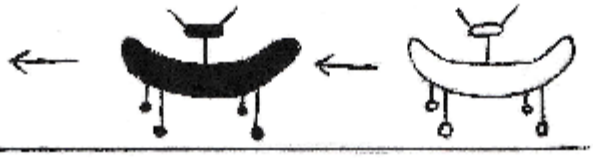
Subject ID #:_____

Please circle TRUE or FALSE for questions 1-4 below.

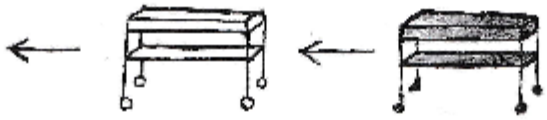
1. The white wheel is in front of the striped wheel. **TRUE / FALSE**



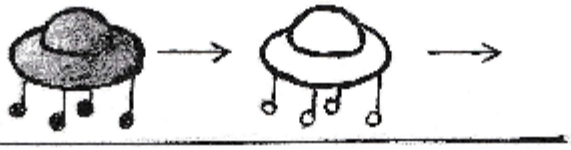
2. The dark blicket is in front of the white blicket. **TRUE / FALSE**



3. The dark cart is in front of the white cart. **TRUE / FALSE**



4. The white widget is in front of the dark widget. **TRUE / FALSE**



On the following page, participants will be instructed: "Please read the following

sentence carefully and then answer the question that appears below.

Next Wednesday's meeting has been moved forward two days.

Which day is the meeting now that it has been rescheduled? (circle the correct answer below)"

Then participants select an answer from a list of seven days (Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday). Finally, participants rate their confidence in their answer on a 5-point scale from 1= not at all confident to 5 = very confident.

Analysis plan. Participants who do not correctly answer all four priming questions will be removed from the analyses. A two-way contingency table will be built with Prime condition (ego-moving vs. object-moving) and Response (Monday vs. Friday) as factors. The critical replication hypothesis will be given by a χ^2 test between prime consistent and inconsistent responses, collapsing across priming condition. In addition, a χ^2 test will be conducted on the control condition in order to determine if there was a bias toward responding Monday or Friday in the absence of any prime.

Known differences from the original. Like the present study, the original study embedded the priming questions and questions about the meeting in a larger survey. However, the measures in the original large survey were different (this task is completed on a clipboard either before or after the next effect).

In the original article, Study 1 had just the spatial ego-moving and spatial object-moving conditions that are included in this replication. The original Study 2 included three additional conditions: control, temporal ego-moving, and temporal time-moving. The original author recommended adding all three conditions. We added the control condition to clarify how the ego and object primes were affecting judgments compared to no priming. Also, the control condition can clarify if participants are highly skewed in answering the "moving forward" dependent variable question as "Friday" or "Monday". If participants show consensus in a single direction, then the manipulations may not be able to overcome this biasing tendency. We did not add the temporal conditions from the original Study 2 to the Study 1 replication. As suggested by the original author, if a null effect is observed, we will not be able to distinguish whether that is particular to spatial priming affecting the outcome, or whether it is a more general lack of temporal priming affecting the outcome in this context. That is, if a time prime does not influence temporal judgments, then there is little chance that a spatial prime could influence temporal judgments. Likewise, if a positive effect is observed, we will not know if that is specific to the spatial priming or also occurs for what might be considered a precondition: time primes influencing time judgments. Following the original, this study is administered as a paper-pencil task. As a result, it cannot be included in the comparison MTurk sample.

3. Availability: A heuristic for judging frequency and probability (Tversky & Kahneman, 1973, Study 3)

Materials and procedure. Participants will read a brief set of instructions about the task:

"The frequency of appearance of letters in the English language was studied. A typical text was selected, and the relative frequency with which various letters of the alphabet appeared

in the first and third positions in words was recorded. Words of less than three letters were excluded from the count.

You will be given several letters of the alphabet, and you will be asked to judge whether these letters appear more often in the first or in the third position, and to estimate the frequency with which they appear in these positions."

Starting on the next page, participants will get a question about each of the letters: K, L, N, R, and V. A sample question is:

"Consider the letter R.
Does R appear more in frequently in (select one)
___the first position?
___the third position?
For every 10 occurrences of R in the first position, how many times will R appear in the third position? _____"

Participants will select either the first or third position option. To estimate the ratio, participants will enter a value into an open-response box. The questions for each letter will appear on its own page, and the order of the pages will be randomized.

Analysis plan. All participants with data will be included in the analysis. We will conduct an analysis strategy very close to the original. Participants who judge the first position to be more frequent for the majority of the letters will get a sign of +, and participants who judge the third position to be more frequent for the majority of the letters will get a sign of -. The number of +s and –s will be used in the sign test.

With our focus on effect size and different response options, we also considering a follow-up analysis strategy with the ratio responses. The following is our present plan for that analysis, but this will undergo more intensive review prior to observing the outcomes of the data collection. The estimated number of occurrence of R in the third position (O3) compared to the first position is the primary variable for analysis. We will recode the data for each response as follows:

If O3 > 10, then SCORE = O3/10 - 1
If O3 = 10, then SCORE = 0
If O3 < 10, then SCORE = 1 - 10/O3

This rescales the ratio estimate to theoretically normalize the distribution and recenters on 0 indicating no difference in frequency estimates. Then, the five ratings will be averaged to create a single index of relative estimation for first versus third letter. It is conceivable that the distributions for these responses will be unusual and require some adjustment in data preparation prior to inferential testing.

Using a *t*-test, the mean rating will be compared against zero indicating no difference in frequency estimates between first and third position. Another possibility is to rescale the distribution for each letter so that 0 indicates the actual difference in frequency of 1st and 3rd position, with positive values indicating overestimation of 1st position and negative values indicating overestimation of 3rd position. The latter would be more definitively an estimate of the magnitude of the misperception.

Finally, as a secondary test, we will examine the effect of order on responses. Is the

overestimation of 1st position more extreme for the first estimate compared to the others? This was not examined in the original research, but it is possible that the effect is dampened with multiple assessments as participants have a meta-experience of realizing that not every letter could be more frequent in the first than third position.

Known differences from original. The original study was on paper. Adapting the task to an online version will permit full randomization of the order of the questions about the letters. With the help of an original author, we changed the instructions slightly, and reworded the part of the paradigm that asks the participant to enter a ratio into an open-response box. These changes were made to make the task easier to understand. Additionally, in the original study, participants were randomly assigned to receive one of two position orders (first position/third position vs. third position/first position); however, in the current study, first position will always be before third position.

4. The Relation between Persistence and Conscientiousness (conceptual replication of De Fruyt, Van De Wiele, & Van Heeringen, 2000)

Materials and procedure. Participants will be given the following instructions: "Below are four strings of letters. Your task is to unscramble the letters to form words. Please type your answer in the spaces provided. Whenever you would like to stop working on this task, press continue to proceed." Four strings of letters are presented, each with an open response box to write an unscrambled word: trypa, aaflt, oneci, acelo. Only the first two anagrams are solvable (answers: party, fatal), and were rated as being moderately difficult in pretesting by Aspinwall and Richter (2002). Below the anagrams will be a button to proceed. Conscientiousness will be measured along with the other Big Five personality traits during the individual differences section after the replicated effects.

Analysis plan. Persistence will be measured as the amount of time spent on the anagrams page before pressing the continue button. The relationship between persistence and conscientiousness will be assessed as a pearson's correlation.

Known differences from original. The current study will examine variation in persistence on this measure across the semester. Also, we seek to conceptually replicate the relation between persistence and conscientiousness using this behavioral measure of persistence. This is not a direct replication as the original work examined self-perception of persistence and a long-form personality measure. No study has examined the relationship between conscientiousness and this brief behavioral task. These differences are likely to reduce the estimated correlation between these constructs.

Past unsolvable anagrams tasks have used different anagrams and different numbers of anagrams. The anagrams above came directly from the list of those used by Aspinwall and Richter (2002). However, given the time constraints of this study, only four anagrams will be used, compared to the longer list used by those researchers, and participants will be given a maximum of four minutes to attempt the anagrams. The anagrams selected were pretested in order to ensure that they produced enough variation in the outcome. Since it is a different version of the task, though, it is difficult to compare the results of this persistence measure to past investigations that have used similar tasks.

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

Materials and procedure. Participants will be randomly assigned to the high-power

condition or the low-power condition. Participants assigned to the high-power condition will receive the following instructions: "Please recall a particular incident in which you had power over another individual or individuals. By power, we mean a situation in which you controlled the ability of another person or persons to get something they wanted, or were in a position to evaluate those individuals. Please describe this situation in which you had power—what happened, how you felt, etc."

Participants assigned to the low-power condition will receive the following instructions: "Please recall a particular incident in which someone else had power over you. By power, we mean a situation in which someone had control over your ability to get something you wanted, or was in a position to evaluate you. Please describe this situation in which you did not have power—what happened, how you felt, etc."

Below each set of instructions there will be a text box that allows for 21 lines of writing. Then, as occurred in the original study, participants will complete two brief filler tasks:

"Please list all of the numbers divisible by 3 in the following table.

3 B 5 8 9 0 B
2 2 6 8 B 8 9
1 7 B 4 4 6 3
3 B B 8 B 9 2"

This task will have an open response box at the end. Then participants will be ask to "Count backward by 3 from the number listed: "360" with 10 open-ended response boxes following.

After these filler tasks, participants will read the following scenario:

"While on a business trip, you and a colleague went to a restaurant that had been recommended by a friend of your colleague. You both dislike your dinner and your colleague sent the following e-mail to the friend that recommended the restaurant. This was the only information that your colleague's friend received about the restaurant: 'About the restaurant, it was marvelous, just marvelous'."

Participants answer "How do you think the colleague's friend will take the comment?" on a 6-point Likert-scale ranging from 1 = *very sarcastic* to 6 = *very sincere*.

Analysis plan. Following the original study, all participants with data on the dependent variable will be included in the analysis. An independent samples *t*-test will be conducted comparing the sarcastic-sincere ratings for participants in the high-power and low-power conditions. As a secondary analysis for this project, we will investigate whether the length of participants' responses to the power prime (as measured by the number of characters in their response) moderates this effect.

Known differences from original. None.

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

Materials and procedure. This task will be completed while standing during the face-to-face portion of the procedure. Participants will be randomly assigned to the heavy or the light clipboard condition. Participants will be instructed to complete the survey on their assigned clipboard. There will be no place to sit, and the experimenter will instruct the participant casually to just fill out the short survey "right here." Participants will first read about a university

committee that has not allowed students to express their opinion about the size of the study abroad grant (taken from Van den Bos, Wilke, & Lind, 1998):

"Read the scenario below carefully. Imagine the following scenario as well as possible, and try to live in the scenario:

You would like to spend 6 months in Amsterdam to conduct research for your Master's thesis. The university you would like to attend is highly recommended. You will work together with highly esteemed professors. All of this offers you good career opportunities. Furthermore, Amsterdam has some other advantages as well: for example, the city, the sights, the culture, and the night life. To pay for your stay and research in Amsterdam, you apply for a grant at "Students Around the World" (SAW). To decide whether they will award you the grant, you will have to appear before the grant committee of SAW.

You appear before the committee. The committee gives you no voice: The committee *does not ask you* to voice your opinion about the amount of money you think you need for your stay and research in Amsterdam.

Now think for a moment about what it means to not be able to voice your opinion about what you need."

Participants will then be asked to answer "How important it was for them that the committee would listen to the opinion of the students?" on a scale ranging from 1 = *not at all* to 7 = *very much*.

Analysis plan. Following the original study, all participants answering the dependent variable will be included in the analysis. The differences between the two groups will be measured using an independent samples *t*-test.

Known differences from original. The study will be administered in English instead of Dutch as was the original. This has led to some slight revisions of the materials through translation. Notably, the destination in the vignette was changed from California to Amsterdam, and the attractions of that site were changed accordingly.

The original clipboard was metal. Three data collection sites for this study have a very similar metal clipboard. The remaining have a plastic clipboard. The clipboards will be weighted identically as the original, and the effects will be examined separately between the metal and plastic clipboards to test whether this makes a difference.

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

Materials and procedure. Participants will read either the communal or agentic description of Marta or Mark, then answer three questions about the person described, and then answer three questions about the room they are sitting in. A male and female version of the descriptions is selected randomly to account for any effects of target gender. The descriptions are identical except for the name and pronouns used.

The communal condition description is:

"Marta/Mark is a nice brown-eyed brunette. Her/His friends say that they can always count on

her/him. They know that Marta/Mark is ready to listen to them and give them advice. She/He is sensitive and eager to help everyone in need. She/He is unselfish and is always interested in giving support to others – she/he is happy when others find her/his help beneficial. She/He is perceived as friendly and caring. She/He is a loyal person, trying not to let anyone down. She/He respects her/his rules but at the same time she's/he's open for others' needs."

The agentic condition description is:

"Marta/Mark is a young fit brunette. She/He is very active both in her/his work and in her/his private life. She/He is competent and efficient as a person. At work, she/he is very creative and full of inspiring ideas. She/He is perceived as very precise in what she/he does, as well as effective and resolute, which is reflected in her/his career. Besides being absorbed by her/his work, she/he is also very keen on sports. Every day, before going to work, she/he runs with her/his dog, while on weekends she/he rides her/his bike for long distances."

After reading the description, participants are asked three questions about the person: (1) "What is your overall evaluation of presented person?" with a response scale from 1 = very negative to 7 = very positive; (2) "Do you find the person likable?" with a response scale from 1 = not at all likable to 7 = very likable; and (3) "How close do you feel to this person?" with a response scale from 1 = very distant to 7 = very close.

After that, participants are told "We are interested in how people perceive different aspects of interiors. Please answer the following questions": (1) "What is the ambient temperature in the room?" with open response in Celsius or Fahrenheit depending on the norm in the data collection location; (2) "What is your opinion of the lighting in this room?" with a response scale from 1= *It's very dim* to 7 = *It's very bright*; and (3) "How spacious is the room?" with a response scale from 1 = *Very cramped* to 7 = *Very spacious.*

Analysis plan. The original study did not exclude extreme data points. However, on the suggestion of an original author, we defined extreme boundaries for data removal. Participants who estimate the temperature as being higher than 95 degrees Fahrenheit (35 degrees Celsius) or lower than 50 degrees Fahrenheit (10 degrees Celsius) will be removed prior to descriptive and inferential analyses. An independent samples *t*-test will test the difference in temperature estimates between the communal and agentic conditions.

Secondary analyses will be conducted with a multivariate model to test the effects of the manipulation when including additional predictors: target gender, participant gender, the actual temperature of the room, and the interaction between target and participant gender.

Known differences from original. The stimulus materials were originally administered in Polish and were translated to English for this study. Additionally, the original study asked the questions about the lab space using a cover story that the school had recently been renovated. Since this cover story might not be applicable at all of the collection sites, we created a new cover story about perceiving interiors with the guidance of the original authors. Also, the original study included an item to rate the warmth/coldness of the target person in the description. At the suggestion of the original authors, this question was dropped to prevent it from influencing the temperature estimate. Finally, the outlier treatment described above was created for this study and was not included in the original design.

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

Materials and procedure. Participants will be randomly assigned to read either strong or weak arguments following these instructions "The following has been prepared by a journalism student for possible publication in the local newspaper. Please read:"

Strong Arguments Condition

[Participant's University name] students should be required to pass a comprehensive examination in their major prior to graduation. Prestigious universities have comprehensive exams to maintain academic excellence. Universities who have instituted comprehensives have seen a reversal in the national trend toward declining standardized test scores. Moreover, graduate and professional schools show a clear preference for undergraduates who have passed a comprehensive exam. Average starting salaries are higher for graduates of schools with comprehensive exams. Also, schools with the exams attract larger and more well-known corporations to recruit students for jobs. The quality of undergraduate teaching has improved at schools with the exams. Furthermore, the state legislature would increase financial support if exams were instituted, allowing a tuition increase to be avoided. Finally, the National Accrediting Board of Higher Education would give the university its highest rating if the exams were instituted. Given these advantages of comprehensive exams, [Participant's University name] should institute them immediately.

Weak Arguments Condition

[Participant's University name] students should be required to pass a comprehensive examination in their major prior to graduation. Adapting the exams would allow the university to be at the forefront of a national trend and thereby increase its reputation throughout the United States and the world. Graduate students have complained that since they have to take comprehensive examinations, undergraduates should have to take them as well. Moreover, by not administering the exams, a tradition dating back to the ancient Greeks was being violated. Parents have written to administrators in favor of the exams. Also, the exams would increase student fear and anxiety enough to promote more studying. The exams would cut costs by eliminating the necessity for other tests that varied with instructor. Furthermore, the exams would allow students to compare their performance with students at other schools. Finally, job prospects might be improved. Given these advantages of comprehensive exams, [Participant's University name] should institute them immediately.

Next, participants read "We would now like to ask you a few questions about what you just read." and respond to five questions on 9-point scales: (1) "To what extent do you think the communication made its point effectively?" responding from 1 = *not at all* to 9 = *complete*; (2) "To what extent did you like the communication?" responding from 1 = *not at all* to 9 = *very much*; (3) "To what extent do you feel the communication was convincing?" responding from 1 = *not at all* to 9 = *very convincing*; (4) "Considering both content and style, how well written was the communication?" responding from 1 = *poorly written* to 9 = *very well written*; and, (5) "Would you judge the recommendation in the preceding message as being …" responding from 1 = *very poor and uncompelling* to 9 = *very good and compelling*.

Analysis plan. The five items will be averaged as an index of argument quality. In the original study, participants who scored in the upper or lower third of need for cognition were

recruited from the available sample for the study and were labeled as being high or low on the trait for analysis. We will use all participants and treat need for cognition as a continuous measure. As such the key analysis will be a general linear model with condition (strong = 1 vs. weak = -1), need for cognition (mean centered), and their interaction predicting argument quality. The key test is the interaction term to show that the effect of the manipulation is moderated by need for cognition.

Known differences from original. The original study specifically recruited participants that were high or low in need for cognition based on earlier testing and treated them as dichotomous groups. We will not use preselection and will treat need for cognition as a continuous variable. A probable consequence of this design change is that changing away from an extreme groups design may result in a weaker overall effect size for moderation by need for cognition. Also, because the prompts are deceptive and depend on the participant being a student, this study will not be included in the MTurk sample.

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

Materials and procedure. Participants will be randomly assigned to either the "best grade" or "worst grade" condition and complete a 5-item academic questionnaire that asks: [1] "When was the last term that you were in school? (e.g. Winter 2013)" with a free response box; [2] "Considering the courses you took in your last term of school, what was your best final grade? (percentage or letter grade)" with a free response box; [3] "Think of the course on which you received your best [worst] final grade last term. How far away does it feel to you?" with a 10-point response scale from 1 = feels like yesterday to 10 = feels far away; [4] "Since it ended, how frequently have you thought about the course on which you received your best [worst] final grade last term? Thinking about the course can include thinking about exams, assignments, or any other aspect of the course." with a 7-point response scale: almost never, very rarely, rarely, sometimes, often, very often, almost all of the time; [5] "How satisfied were you with your grade in this course?" with a 10-point response scale from 1 = very satisfied to 10 = not at all satisfied. Participants also respond on a 7-point Likert scale ranging from 1 (not very true of me) to 7 (very true of me) to the following statement: "I have high self-esteem." (SISE; Robins, Hendin, & Trzesniewski, 2001).

Analysis plan. First, we will conduct stepwise regression analyses to examine the variables of interest. Subjective distance is the dependent variable. All continuous predictor variables will be centered. In the first step, we will enter a variable that indicates actual time, which is calculated from the number of months since the participant's course ended. In the second step, grade condition (best vs. worst grade) and self-esteem are entered simultaneously. In the last step, the interaction between grade condition (best vs. worst grade) and self-esteem will be entered. Follow-up tests will clarify the nature of the interaction effect.

Known differences from original. We will use the single-item measure of self-esteem (SISE; Robins, Hendin, & Trzesniewski, 2001), instead of the Rosenberg Self-Esteem Scale (1965). Also, self-esteem will be analyzed continuously. Unlike the original study, which randomly presented the self-esteem measure before or after the academic questionnaire, the current study will always present the Single-Item Self-Esteem measures after the academic questionnaire.

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

Materials and procedure. Participants will be randomly assigned to the some or most condition and answer whether each of five statements are right or wrong:

1. Most [Some] women are better off at home taking care of the children.
2. Men are more emotionally suited for politics than are most [some] women.
3. The best job for most [some] women is something like cook, nurse, or teacher.
4. Most [Some] women need a man to protect them.
5. Most [Some] women are not really smart.

Next, participants will answer three filler items that provide a plausible alternative basis for the purpose of the dependent variable:

"Are you at all familiar with the building industry (trainings, relatives, ...)?" (1 = not at all familiar to 4 = familiar)

"Given your interests, how likely is it that you should be working in the building industry in the future?" (1 = very unlikely to 5 = very likely)

"Have you ever taken a recruiting decision in your life?" (1 = never to 4 = very often)

Next, participants will read a scenario and assess the suitability of a highly-paid, technical, negotiating position for a building company for male or female employees:

"Imagine that you are the manager of a small (45-person) cement manufacturing company based in New Jersey. Last year was a particularly good one, and after you invested in increasing the output capacity of your plant, you decide that it would be very fruitful if you could find clients in other states to increase your business. Because you cannot spend too much time away from the plant, you decide to appoint someone to go around to prospective clients and negotiate contracts. This is a highly specialized market, and the job will mostly consist in going from one building site to another, establishing contacts with foremen and building contractors. It is also a highly competitive market, so bargaining may at some points be harsh. Finally, it's a very technical market, and a representative that did not exude confidence in their technical skills would not be taken seriously by potential clients. Realizing how useful such a help would be for you, you decide to give the person chosen one of the top-five salaries in your company."

Finally, participants answer two questions as the dependent measures:

"Do you feel that this job is better suited for one gender rather than the other?" (-3 = yes, much better for women; 0 = no I do not feel this way at all; +3 = yes, much better for men)

"In general, to what extent do you agree with the statement that women are just as able as men to do any kind of job?" (-3 = disagree strongly to +3 = agree strongly)

Analysis plan. Following the original study, all participants with data in the first dependent item will be included in the analysis. The primary effect of interest for this replication is the condition (some/most statements) x gender (male/female) interaction, with an expected difference between conditions among males only. Participants' responses to the first dependent

measure question will be subjected to a 2x2 factorial ANOVA. The second dependent measure question is available for secondary analysis.

Known differences from original. The original study recruited participants on a college campus, whereas this study will recruit students from department participant pools. Also, for simplicity of design, this study will drop the base rate condition (participants who did not see any sexist statements). Follow-up note (March 8, 2015): While not in the pre-registered plan, the primary effect of interest in the original article was the main effect of credentials condition. That main effect is also reported in the main text.

Supplementary Information: Tables and Figures

See Supplementary Tables Spreadsheet for these materials.

Supplement: Detailed Analysis Report for Times of Semester Effects

1. Stroop task (Stroop, 1935). The unconditional model revealed that inter-site variation accounted for .6% of the variance in Stroop effect scores. The full mixed effects model contained Time of Semester as a fixed effect and a random intercept and random slope of Time of Semester by site. Correlations between the random effects revealed that our model was overparameterized, so the random slope was dropped from the random effects (leaving a random intercept of site). The final model was compared to a model without Time of Semester as a fixed effect. The addition of the Time of Semester fixed effect did not reliably improve model fit, $\chi^2(1, N = 2660) = 3.31$, $p = .069$, $\Phi = .04$. This failed to provide evidence for time of semester variation for this effect. Following up on this, we constructed a simple linear model predicting Stroop score from Time of Semester. In this model, Time of Semester reliably predicted Stroop performance, with participants showing a stronger Stroop effect as the semester wore on, $F(1, 2658) = 5.01$, $p = .025$, $\eta_p^2 = .002$, 95% CI = [0, .007].

2. Metaphoric structuring (Boroditsky, 2000). For this effect, we constructed generalized linear mixed effects models using the binomial family. Using a binomial constant of $\pi^2/3$, the unconditional model revealed that inter-site variation accounted for less than .01% of the variance of our dependent measure (whether or not the participant responded in line with the priming manipulation). The full mixed effects model contained Time of Semester as a fixed effect and a random intercept and random slope of time of semester by site. Correlations between the random effects revealed that our model was overparameterized, so the random slope of Time of Semester was dropped from the random effects (leaving a random intercept of site). The final model was compared to a model without the Time of Semester fixed effect. The addition of Time of Semester improved the model fit, $\chi^2(1, N = 1332) = 4.48$, $p = .034$, $\Phi = .06$. To better understand this time of semester variation, we conducted a logistic regression predicting prime consistent responding from Time of Semester, but found no effect $\chi^2(1, N = 1332) = 0.010$, $p = .920$, $\Phi = .003$. However, comparing participants from the first 80% of the semester to the last 20% of the semester, we observed that the effect was strongest at the very end of the semester compared to the rest of the semester ($d = .36$ compared to $d = .24$).

3. Availability heuristic (Tversky & Kahneman, 1973). For this effect, we constructed generalized linear mixed effects models using the binomial family. Using a binomial constant of $\pi^2/3$, the unconditional model revealed that inter-site variation accounted for less than .01% of the variance of our dependent measure (whether a majority of letters were predicted to occur more frequently in the first position). The full mixed effects model contained Time of Semester

as a fixed effect and a random intercept and random slope of time of semester by site. Correlations between the random effects revealed that our model was overparameterized, so the random slope of Time of Semester was dropped from the random effects (leaving a random intercept of site). The final model was compared to a model without the Time of Semester fixed effect. The addition of Time of Semester did not reliably improve the model fit, $\chi^2(4, N = 2497) = 1.45$, $p = .228$. This failed to provide evidence for time of semester variation for this effect.

4. Persistence and Conscientiousness (De Fruyt, Van De Wiele, & Van Heeringen, 2000). For this effect, we created linear mixed effects models predicting persistence (measured as the number of seconds spent on the unsolvable anagrams task) from conscientiousness and time of semester. The unconditional model revealed that inter-site variation accounted for 5% of the variance in persistence scores. The full mixed effects model contained the Conscientiousness × Time of Semester interaction as a fixed effect and a random intercept and random slope of this interaction by site. Correlations between the random effects revealed that our model was overparameterized, so the interaction term and its components were dropped from the random effects (leaving a random intercept of site). The final model was compared to a model without the Time of Semester interaction as a fixed effect. The addition of the Time of Semester interaction did not reliably improve the model fit, $\chi^2(2, N = 2624) = 4.63$, $p = .099$. This failed to provide conclusive evidence for time of semester variation for this effect.

5. Power and perspective (Galinsky et al., 2006). The unconditional model revealed that inter-site variation accounted for 0.9% of the variance in sarcasm ratings. The full mixed effects model contained the Power Condition × Time of Semester interaction as a fixed effect and a random intercept and random slope of this interaction by site. Correlations between the random effects revealed that our model was overparameterized, so the interaction term and its components were dropped from the random effects (leaving a random intercept of site). The final model was compared to a model without the Time of Semester interaction as a fixed effect. The addition of the Time of Semester interaction did not reliably improve the model fit, $\chi^2(2, N = 2385) = .70$, $p = .699$. This failed to provide evidence for time of semester variation for this effect.

6. Weight embodiment (Jostmann et al., 2009). The unconditional model revealed that inter-site variation accounted for 0.6% of the variance in importance ratings for the presented issue. The full mixed effects model contained the Clipboard Weight × Time of Semester interaction as a fixed effect and a random intercept and random slope of this interaction by site. Correlations between the random effects revealed that our model was overparameterized, so the interaction term and its components were dropped from the random effects (leaving a random intercept of site). We compared the final model to a model without the Time of Semester interaction as a fixed effect. The addition of the Time of Semester improved the model fit, $\chi^2(2, N = 2279) = 30.46$, $p < .001$. To follow up on this finding, we constructed a linear model predicting importance ratings from the Clipboard Weight × Time of Semester interaction, but found no effect, $F(1, 2275) = 0.31$, $p = .578$, $r = .01$. Upon further examination, the full model suffered from multicollinearity between our fixed effects, likely producing a spurious model improvement. Observing just the full model, the Clipboard Weight × Time of Semester fixed interaction was not reliable, $t(2275) = -0.57$, $p = .569$.

7. Warmth perceptions (Szymkow et al., 2013). The unconditional model revealed that inter-site variation accounted for 22% of the variance in temperature estimates. This indicates that that labs varied in their average temperature. The full mixed effects model contained the Temperature Condition × Time of Semester interaction as a fixed effect and a

random intercept and random slope of this interaction by site. Correlations between the random effects revealed that our model was overparameterized, so the interaction term and Temperature Condition were dropped from the random effects (leaving a random intercept and slope of Time of Semester by site). The final model was compared to a model without the Time of Semester interaction as a fixed effect. The addition of the Time of Semester interaction reliably improved the model fit, $\chi^2(2, N = 2544) = 6.04$, $p = .049$. This failed to provide evidence for time of semester variation for this effect. To unpack this finding, we constructed a linear model predicting temperature estimate from the Temperature Condition × Time of Semester interaction, controlling for actual temperature in the lab. This model did not reveal a reliable Condition × Time of Semester interaction, but did reveal a weak main effect of Time of Semester, such that temperature estimates overall tended to decline over the course of the semester, $F(1, 2540) = 2.95$, $p = .086$, $\eta_p^2 = .001$, 95% CI = [0, .005]. The most obvious explanation for the slight model improvement is declining temperature in North America during the Fall semester.

8. Elaboration Likelihood (Cacioppo, Petty, & Morris, 1983). The unconditional model revealed that inter-site variation accounted for 1.1% of the variance of our dependent measure (rating of argument quality). The full mixed effects model contained the Argument Condition (strong or weak arguments) × Need for Cognition (centered) × Time of Semester three-way interaction as a fixed effect and a random intercept and random slope of the three-way interaction by site. Correlations between the random effects revealed that our model was overparameterized, so the interaction term and its components were dropped from the random effects (leaving a random intercept of site). The final model was compared to a model without the Time of Semester three way interaction as a fixed effect. The addition of the Time of Semester interaction did not reliably improve the model fit, $\chi^2(4, N = 2365) = 2.02$, $p = .732$. This failed to provide evidence for time of semester variation for this effect.

9. Self-esteem and subjective distance (Ross, & Wilson, 2002). The unconditional model revealed that inter-site variation accounted for 0.9% of the variance of our dependent measure (felt distance from the recalled course). The full mixed effects model contained the actual months since the recalled course and the Grade Condition (Best or Worst grade recalled from last term) × Self-Esteem × Time of Semester three-way interaction as a fixed effect and a random intercept, random slope of months since the course, and random slope of the threeway interaction by site. Correlations between the random effects revealed that our model was overparameterized, so the interaction term and its components were dropped from the random effects, as well as the random slope of months since the course (leaving a random intercept of site). The final model was compared to a model without the Time of Semester three way interaction as a fixed effect. The addition of the Time of Semester interaction did not reliably improve the model fit, $\chi^2(4, N = 2562) = .54$, $p = .969$. This failed to provide evidence for time of semester variation for this effect.

10. Credentials and prejudice (Monin & Miller, 2001). The unconditional model revealed that inter-site variation accounted for 0.4% of the variance of our dependent measure (whether or not participants believed a job was more appropriate for one gender). The full mixed effects model contained the Credential Condition × Gender × Time of Semester interaction as a fixed effect and a random intercept and random slope of this interaction by site. Correlations between the random effects revealed that our model was overparameterized, so the interaction term and its components were dropped from the random effects (leaving a random intercept of site). The final model was compared to a model without the Time of Semester three way interaction as a fixed effect. The addition of the Time of Semester interaction did not reliably

improve the model fit, $\chi^2(4, N = 2571) = 4.90$, $p = .298$. This failed to provide evidence for time of semester variation for this effect.

Elaboration Likelihood (argument quality main effect) To test time of semester variation for the observed main effect of argument condition, we compared a linear mixed effects model containing the Argument Condition x Time of Semester interaction as fixed effects, with a random intercept of site, to a model containing only the Argument Condition fixed effect. The addition of Time of Semester did not reliably improve the model, $\chi^2(2, N = 2429) = 0.22$, $p = .896$.

Self-esteem and subjective distance (grade condition main effect) To test time of semester variation for the observed main effect of grade condition, we compared a linear mixed effects model containing actual months since the recalled course and the Grade Condition x Time of Semester interaction as fixed effects, with a random intercept of site, to a model containing only the Actual Months and Grade Condition fixed effects. The addition of Time of Semester did not reliably improve the model, $\chi^2(2, N = 2562) = 0.32$, $p = .851$.

Credentials and prejudice (credentialing main effect). To test time of semester variation for the observed main effect of credentialing condition, we compared a linear mixed effects model containing the Credentials Condition x Time of Semester interaction as a fixed effect, with a random intercept of site, to a model containing only the Credentials Condition fixed effect. The addition of Time of Semester did not reliably improve the model, $\chi^2(2, N = 2642) = 2.15$, $p = .341$.

Supplemental Materials: Additional Moderator Analyses

To determine the possible influence of the data quality indicators and individual differences that varied over the semester on the effects observed, we conducted moderator analyses for each of the individual differences and data quality indicators as a moderator of each of the replicated effects. There were a few instances where effects were reliably ($p < .05$) moderated by these factors. For Availability Heuristic, those who failed the attention check ($d = .08$), reported less effort ($d = -.09$), and were male showed the effect more strongly ($d = .14$). For Elaboration Likelihood, women were more affected by argument strength than men ($\eta_p^2 = .002$). Finally, the main effect in Self-Esteem and Subjective Distance was moderated by both mood ($\eta_p^2 = .002$) and stress ($\eta_p^2 = .012$). For those in the Best Grade condition, mood was positively related to feeling close to the recalled class. Stress interacting with both conditions, but in different ways, relating to feeling closer to the Worst Grade class and feeling further from the Best Grade class. Although one could craft plausible theoretical accounts for these moderation effects, we recommend caution in interpreting these findings for two reasons. These analyses are all exploratory and not part of the preregistered analysis plan. Also, they are the result of 42 possible statistical tests. In fact, finding 6 significant findings out of 42 is not reliably different from what one would expect to find given false positive rates for using an alpha level of .05, $\chi^2(1, N = 42) = .182$, $p = .670$.