PNAS
www.pnas.org

Supplementary Information for

Generic language in scientific communication

Jasmine M. DeJesus, Maureen A. Callanan, Graciela Solis, and Susan A. Gelman

Jasmine M. DeJesus
Email: jmdejes2@uncg.edu

This PDF file includes:

Supplemental text
Fig. S1 to S3
Tables S1 to S9
Supplemental references

Other supplementary materials for this manuscript include the following:

Link to raw data files:
https://osf.io/v8nqe/?view_only=900f8247b4d34c568207e15835c99083

Data in all studies were analyzed in R (1). Data files and analysis code are available on the Open Science Framework (OSF): https://osf.io/v8nqe/?view_only=900f8247b4d34c568207e15835c99083

Study 1 supplemental method

Article inclusion. All articles were published in 2015 and 2016, the most recent complete years available when initiating the study, in *Biological Psychology* (BP), *Cognition* (Cog), *Cognitive Psychology* (CP), *Developmental Science* (DS), the *International Journal of Psychophysiology* (IJP), the *Journal of Abnormal Psychology* (J Ab Psy), the *Journal of Consulting and Clinical Psychology* (JCCP), the *Journal of Experimental Child Psychology* (JECP), the *Journal of Experimental Social Psychology* (JESP), the *Journal of Memory and Language* (JML), and the Psychological and Cognitive Sciences and Social Sciences sections of the *Proceedings of the National Academies of Science* (PNAS); see Table S1 for details about the requirements of each journal. Articles were included in the analysis if they provided research with human participants. Therefore, articles were excluded if they included only non-human animals, model simulations, meta-analyses, corrections, or material such as commentaries, editorials, or discussion ($N = 92$), or if they did not provide research highlights or summaries ($N = 121$); see Table S2 for a breakdown of exclusions. The remaining 1,149 articles were included for more detailed coding.

Coding. Titles, highlights, and abstracts were assembled into spreadsheets with one row for each sentence or sentence fragment (such as bullet points) in each article; rows are referred to henceforth as "lines." Coding took place in two phases. First, the four authors assessed whether each line (including titles, highlights, and abstracts) described the results of the study. Lines categorized as "not results" described the prior research, the research question, the study hypotheses, or the methods. Lines were also categorized as "not results" if they were concluding sentences about future work that is needed, a description of how issues were or will be discussed in the article, or a vague statement about the field (e.g., "Future directions for investigating face processing development in biracial populations are discussed."). Second, the four authors assessed whether each line included generic language using the following decision tree:

1. If the line is not a complete sentence, then it is uncodable.
2. If the sentence is exclusively in the past tense, then it is not generic.
3. If the sentence is not exclusively in the past tense, and if it makes a broad claim that refers to categories (e.g., "children") or abstract concepts (e.g., "parental warmth") instead of specific exemplars (e.g., "the children tested in these experiments" or "the warmth of parents in this study"), then it is generic. A key question posed to coders was: Are the authors extending their findings generally to members of the category or instances of the phenomenon tested, or are they describing their results in terms of the specific participants or study?
4. If the sentence is generic, the next step is to decide whether the generic is bare (i.e., has no additional information to qualify or frame the results), hedged, or framed.

To promote consistent coding, a set of phrases was designated as indicators of either hedged or framed generics prior to coding or through discussions of disagreements. Hedge phrases included: Suggest, support, imply, consistent with, in line with, conform with, perhaps, may, might, can, could, usually, often, call into question, cast doubt, propose, argue against. Frame phrases included: Report, exhibit, demonstrate, show, confirm, reveal, establish, indicate, highlight, emphasize, implicate, illustrate, provide evidence, negate.

Reliability. For both phases of coding (determining which lines reported results, and generic language coding), four coders (the authors) coded an overlapping set of roughly 20% of the articles ($N = 259$). Reliability was calculated for each pair of coders, kappas > .90, with an average of .94 across coders for results-coding (Step 1), and kappas > .66 with an average of .76 across coders for generics-coding (Step 2). Disagreements were resolved by discussion. The remaining articles were divided equally among the four coders for independent coding.

To further check the independent coding, coders exchanged the independent coding and checked for agreement. Changes were made to 375 sentences (an additional 56 sentences were discussed but not changed) out of 14,878 total units and 8,992 units that pertained to results. These changes affected 234 articles out of 1,149.

Article entry. Several features of each article were entered by a group of trained student research assistants; see Table S3 for descriptive statistics for those features divided by journal. Coders were instructed to look in the participants section of the article and also to look for any available tables of participant demographics in the main text. For each article, the following variables were analyzed:

1. The number of participants included in the study (minus excluded participants if reported).
2. The country in which participants were recruited: Not specified, in the United States only, or U.S. and beyond. "U.S. and beyond" included studies with participants in the United States *and* another country, or with participants from one or more countries outside of the U.S. This was not assumed from the author affiliations but was included if explicitly stated, or if an ethics committee from a specific university was mentioned.
3. Participant race: Unspecified (i.e., no information about the racial background of participants was provided) or specified (ranging from "predominantly White" to specific breakdowns of participants by race/ethnicity).
4. Participant socioeconomic status (i.e., no information about the socioeconomic background of participants was provided) or specified in some way. This was a broad category and could include reporting indicators such as educational history (other than reporting that participants were college students), neighborhood resources, or family income.
5. Participant language: Unspecified (i.e., no information about the language background of participants was provided) or specified (i.e., the authors mentioned that speaking a particular language was an inclusion criterion for the study, or authors recruited people who spoke in specific languages).

A set of roughly 20% of the articles ($N = 233$) were coded by a second coder to assess reliability. To assess reliability for the number of participants reported in each article, an intraclass correlation was computed to compare the two sets (two-way, agreement), ICC = .99 (CI: .99, .994). For all other categories, reliability was assessed using Cohen's kappa to compare the two sets: Test country = .74; participant race = .87; participant socioeconomic status = .71; participant language = .77. Students were instructed to write any comments indicating difficulty or confusion; those responses were reviewed by a more expert coder (with a Ph.D. in Psychology). Disagreements were resolved by the expert coder.

Study 2 supplemental method

Participants. Participants were recruited by Amazon's Mechanical Turk with the criterion that they were located in the United States, were native English speakers, and had "been granted the Mechanical Turk Masters Qualification" by Amazon, meaning that workers "have consistently demonstrated a high degree of success in performing a wide range of HITs across a large number of Requesters" (see Supplemental Table 4 for additional participant characteristics). Participants were paid $1.50 for completing the study. The study was conducted in February 2018.

Our target sample size for each of the four test questions was 100 participants. Over two rounds, a total of 551 participants began the survey. The vast majority of participants were in the first round ($N = 506$); the survey was offered to an additional 45 participants to reach 100 participants per question.

Participants were excluded from analysis if they completed fewer than 48 of the 60 items ($N = 128$). This criterion was determined by examining a histogram of the number of items completed by participants and selecting a natural "break" that permitted inclusion of as many participants as possible while excluding participants who did not complete a large portion of the study (see Figure S1). Among the excluded participants, 63 completed 0 items, 53 completed 1-11 items, and 12 completed 22-47 items. For the 63 excluded participants who did not complete any items, we cannot determine which question they were assigned to receive. For the 65 excluded participants who completed at least one item but did not reach our inclusion criterion, 15 were excluded from the importance question, 30 were excluded from the generalizability question, 11 were excluded from the sample size question, and 9 were excluded from the diversity question.

An additional 7 participants were excluded if they reported that they were not native English speakers, given that generics are expressed differently in different languages, and non-native speakers of English may have difficulty with the generic-non-generic distinction in English (2).

A total of 416 participants were included in our analyses. All included participants completed at least 57 items out of 60 (i.e., all participants who completed at least 48 items completed nearly the entire study). Because only 8 excluded participants answered

any demographic questions, we cannot determine whether included and excluded participants differed on any demographic variables.

Materials. We selected 60 titles from the Study 1 corpus that were unanimously coded as bare generics by the four coders during the reliability phase, and that had relative higher readability scores (i.e., a lower U.S. grade level would be needed to comprehend the text based on the Flesch Kincaid Grade Level: https://www.online-utility.org/english/readability_test_and_improve.jsp). Ten titles each were selected from five different content areas of psychology (biological, clinical, cognitive, developmental, social) and *PNAS* (60 total). Hedged, framed, and non-generic versions of each title were created from the bare generic version to control for article content across participants. Titles were described as "a brief summary of different research projects" and participants were randomly assigned to complete one of four test questions for each summary:

Procedure. Participants completed a Qualtrics survey accessed through Amazon's Mechanical Turk worker interface. Participants were randomly assigned to one of four questions, which asked them to evaluate the importance of the finding, the generalizability of the finding (measured as the percentage of people to whom participants thought the finding applies), the sample size of the study, and the extent to which the finding applies to diverse participants. They were also randomly assigned to one of four survey versions within each question. Each version was created so that participants saw each summary only once (as either non-generic, bare, framed, or hedged) with 10 summaries per domain, and so that each summary could be presented in each language type across participants. Within each survey version, summaries were presented in random order.

Importance. Survey instructions were the following:

> In this survey, you will see a series of brief summaries of different research projects. These summaries are based on the titles of actual published research papers.
>
> For each project, please read the brief summary and then give your best guess of how important the research project is on a scale ranging from 1 (not important at all) to 7 (extremely important).
>
> You might see words that you do not recognize or jargon. Please give your first impression of each summary. Read the sentence once or twice and give the first answer that comes to mind.
>
> You will see 60 summaries in this survey.

Generalizability. Survey instructions were the following:

> In this survey, you will see a series of brief summaries of different research projects. These summaries are based on the titles of actual published research papers.

> Every project is based on a sample of participants. Each sample includes a set number of people who completed the research project and each sample has unique characteristics. In this survey, we want to know what percentage of those in the world today would show the effect described in the research summary. For each summary, please give your best guess based on the information provided.
>
> You might see words that you do not recognize or jargon. Please give your first impression of each summary. Read the sentence once or twice and give the first answer that comes to mind.
>
> You will see 60 summaries in this survey.
>
> [Participants were asked to enter a number from 0% to 100% to represent the percentage of people to whom the finding applies.]

Sample size. Survey instructions were the following:

> In this survey, you will see a series of brief summaries of different research projects. These summaries are based on the titles of actual published research papers.
>
> For each project, please read the brief summary and then give your best guess of how many people participated in the research project from the options provided.
>
> You might see words that you do not recognize or jargon. Please give your first impression of each summary. Read the sentence once or twice and give the first answer that comes to mind.
>
> You will see 60 summaries in this survey.
>
> [Participants were asked to report the number of participants as one of seven ranges: (1) 1-10 people, (2) 11-50 people, (3) 51-100 people, (4) 101-250 people, (5) 251-500 people, (6) 501-1000 people, (7) 1001 people or more.]

Diversity. Survey instructions were the following:

> In this survey, you will see a series of brief summaries of different research projects. These summaries are based on the titles of actual published research papers.
>
> Every project is based on a sample of participants. Each sample includes a set number of people who completed the research project and each sample has unique characteristics. In this survey, we want to know how likely you think it is that the effect described in the research project would extend to people from diverse backgrounds (for example, differing in nationality, race, ethnicity, socioeconomic status, etc.).

For each summary, please give your best guess based on the information provided on a scale ranging from 1 (not likely at all) to 7 (extremely likely).

You might see words that you do not recognize or jargon. Please give your first impression of each summary. Read the sentence once or twice and give the first answer that comes to mind.

You will see 60 summaries in this survey.

Study 2 results. See Supplemental Table 5. Overall, we found that study summaries described using generics were rated as more important (for all types of generics, *p*s < .02), more generalizable (for framed generics, $p = .023$), and having larger sample sizes (for hedged generics, $p = .038$) compared to summaries described using non-generic language. No significant differences were observed when participants were asked to report whether the finding would extend to people from more diverse backgrounds.

The content area of the summaries strongly influenced participants' judgments on all four test questions. Compared to studies drawn from PNAS, studies in biological psychology journals were rated as having larger samples ($p < .001$). Studies drawn from clinical psychology journals tended to be rated as more important ($p < .001$), less generalizable ($p = .017$), including a larger sample ($p < .001$), and extending to less diverse samples of participants ($p < .001$). Studies drawn from cognitive psychology journals tended to be rated as less important ($p < .001$), including a smaller sample ($p < .001$) and extending to less diverse samples of participants ($p < .001$). Studies drawn from developmental psychology journals tended to be rated as including a smaller sample ($p < .001$). Studies drawn from social psychology journals tended to be rated as less important ($p < .001$), less generalizable ($p = .004$), including a larger sample ($p = .02$), and extending to less diverse samples of participants ($p < .001$).

Correlations across questions. We tested for correlations across questions to examine whether participants who responded to different questions provided similar ratings for each summary. To do so, we created an index for each summary by averaging participants' scores for each question across generic or non-generic forms. Two significant correlations were observed: The generalize and diversity questions were positively correlated, $r(60) = .68$, $p < .001$. The higher the percentage (from 0% to 100%) to whom participants expected the findings to apply, the more participants expected the finding to extend to diverse groups of people. In addition, importance and number of participants were positively correlated, $r(60) = .33$, $p = .010$. The more important participants rated the summary, the more participants they expected were in the study. No other correlations were significant (generalize-n: $r(60) = .22$, $p = .088$; diversity-n: $r(60) = .07$, $p = .602$; generalize-importance: $r(60) = .18$, $p = .159$; importance-diversity: $r(60) = .24$, $p = .068$).

Studies 3a-3d supplemental method

Participants. See Table S6. Participants were recruited by Amazon's Mechanical Turk with the criterion that they were located in the United States, native English speakers, had not completed a previous version of this study, and had either been granted the Masters Qualification (Study 3a, $n = 74$) or had completed at least 100 HITs with a 95% approval rating (Studies 3b, 3d, $n = 264$ and 299, respectively). Participants were paid $1.50 for completing the study. These studies were approved under the same IRB protocol as Study 2. Our target sample size for each study and test question was 100 participants. Studies were conducted in February and March 2019.

An additional sample of participants was recruited from the University of Michigan introductory psychology subject pool (Study 3c, $n = 118$). Participants received partial course credit for participating in the study.

Based on the criteria established in Study 2, participants were excluded from the analysis if they completed fewer than 80% of the study items. 70 participants were excluded on this basis (3a = 57, 3b = 13, 3c = 0, 3d = 0). An additional 5 participants were excluded if they reported that they were not native English speakers (3a = 0, 3b = 3, 3c = 1, 3d = 1). Although we set U.S. location as recruitment criteria, 2 participants in Study 3d who were native English speakers reported that they were not currently in the U.S.; we included the data for those participants.

Materials. For Studies 3 and 4, summaries focused on groups of people (e.g., "bilinguals") rather than abstract concepts (e.g., "statistical learning"), to increase readability and provide a direct contrast between participant sample and abstract category to which the sample belonged. As a first step, potential summaries were generated by each of the four Study 1 coders by examining the sentences they coded that described study results (including titles, highlights, or abstracts). Each coder was instructed to look for sentences where (a) a kind of person was in the subject position (not object or possessive), (b) the kind of person was a broad category (e.g., "children" but not "participants"), and (c) the sentence had as little jargon as possible. From that list, coders were then instructed to make any needed changes to simplify the language and create a bare generic (if the sentence was not already in that format). Coders generated a list of 81 sentences, which were then rated by a group of MTurk workers ($n = 46$) for readability. Participants were asked, "In your opinion, how clear and easy to understand was that summary?" and rated each sentence on a scale from 1 (least clear and hardest to understand) to 7 (most clear and easiest to understand). Sentences were selected for subsequent studies based on mean rating (i.e., sentences with a rating of at least 4.5 out of 7) and to include a variety of noun phrases (i.e., to avoid most sentences being about "children" or "people"). The final list included 36 sentences with a mean readability rating of 5.40 (out of 7).

Study 3a procedure. Participants completed a Qualtrics survey accessed through Amazon's Mechanical Turk worker interface. Participants were shown 36 summaries of research findings, each described using one of three language forms: (e.g., "People with dysphoria **are** less sensitive to positive information in the environment."), a non-generic expressed with simple past tense, as was done in Study 2 (e.g., "People with dysphoria

were less sensitive to positive information in the environment."), or a non-generic with multiple cues (past tense, qualifier, and "some"; e.g., "**Some** people with dysphoria **were** less sensitive to positive information in the environment, **under certain circumstances**." Emphases added.). Qualifiers included: "at times", "in certain situations", "in some cases", "some of the time", and "under certain circumstances". Each version was created so that participants saw each summary only once (as either bare generic, past-tense non-generic, or multi-cue non-generic). Each participant received 12 summaries in each of the three forms. A given summary was presented in only one form per participant, but in different forms across participants. Within each survey version, summaries were presented in random order.

For each summary, participants were asked four questions (with order counterbalanced across participants) and provided ratings on a scale of 1 (not at all) to 7 (to a great extent):

- How important is this finding?
- How much would you want to draw conclusions from this finding?
- How much would this finding generalize within the United States?
- How much would this finding generalize outside the United States?

Study 3a results. See Table S7. Across all test questions, participants rated the bare generics (the reference category, $M = 4.52$, 95% $CI = 4.32, 4.72$) more highly than the multi-cue non-generics ($M = 3.85$, 95% $CI = 3.59, 4.11$; $b = -0.77$, $SE = 0.09$, $z = -8.92$, $p < .001$), but not differently from the past-tense non-generics ($M = 4.54$, 95% $CI = 4.34, 4.74$; $b = 0.06$, $SE = 0.08$, $z = 0.66$, $p = .511$).

Compared to the conclude question (the reference category, $M = 4.09$, 95% $CI = 3.89, 4.30$), participants provided higher ratings when asked if the findings were likely to generalize within the U.S. ($M = 4.55$, 95% $CI = 4.30, 4.81$; $b = 0.51$, SE = 0.08, $z = 6.06$, $p < .001$) and outside the U.S. ($M = 4.42$, 95% $CI = 4.15, 4.68$; $b = 0.37$, SE = 0.08, $z = 4.44$, $p < .001$). No significant interaction between generic language and question was observed ($p = .375$).

Studies 3b and 3c procedure. Participants completed a Qualtrics survey. Participants were shown 36 summaries of research findings, each described with a bare generic, a past-tense non-generic, or a multi-cue non-generic, as in Study 3a. As in Study 3a, each participant received 12 summaries in each of the three forms. A given summary was presented in only one form per participant, but in different forms across participants. Each participant was asked to provide only one type of rating (randomly assigned across participants) on a 1 to 7 scale per summary: "How important is this finding" or "How much would you want to draw conclusions from this finding?" Participants were recruited through either Amazon Mechanical Turk (importance: $n = 135$; conclude: $n = 129$) or the University of Michigan undergraduate study pool (importance: $n = 60$; conclude: $n = 58$).

Study 3b results. See Table S7. A significant interaction between generic language and question was observed ($p < .001$); therefore separate models were run for each question. For the importance question, MTurk participants rated the bare generics ($M = 4.53$, 95% $CI = 4.37, 4.68$) more highly than the multi-cue non-generics ($M = 4.11$, 95% $CI = 3.94, 4.28$; $b = -0.54$, $SE = 0.06$, $z = -8.40$, $p < .001$), but not differently from the past-tense non-generics ($M = 4.56$, 95% $CI = 4.42, 4.69$; $b = 0.04$, $SE = 0.06$, $z = 0.55$, $p = .585$). Similarly, for the conclude question, MTurk participants rated the bare generics ($M = 4.38$, 95% $CI = 4.21, 4.55$) more highly than the multi-cue non-generics ($M = 3.67$, 95% $CI = 3.48, 3.87$; $b = -0.88$, $SE = 0.07$, $z = -13.01$, $p < .001$), but not differently from the past-tense non-generics ($M = 4.42$, 95% $CI = 4.26, 4.58$; $b = 0.03$, $SE = 0.07$, $z = 0.44$, $p = .661$).

Study 3c results. See Table S7. Undergraduate student participants rated the bare generics ($M = 4.60$, 95% $CI = 4.46, 4.74$) more highly than the multi-cue non-generics ($M = 3.98$, 95% $CI = 3.81, 4.14$; $b = -0.92$, $SE = 0.10$, $z = -8.88$, $p < .001$), but not differently from the past-tense non-generics ($M = 4.59$, 95% $CI = 4.46, 4.72$; $b = -0.02$, $SE = 0.10$, $z = 0.16$, $p = .874$). No significant difference between question was observed (importance: $M = 4.49$, 95% $CI = 4.32, 4.66$; conclude: $M = 4.27$, 95% $CI = 4.12, 4.43$; $b = 0.25$, $SE = 0.18$, $z = 1.37$, $p = .172$). No significant interaction between generic language and question was observed ($p = .177$).

Studies 3b vs. 3c. See Table S7. In order to examine whether participant recruitment method (MTurk vs. student) affected the results, we conducted an additional analysis including both Studies 3b and 3c, adding participant recruitment method (Mturk vs. student) as a factor in the model. This analysis revealed no difference between the MTurk and student sample for either test question (ps $> .4$) and no interaction between generic language and participant group ($p = .645$).

Study 3d procedure. Participants completed a Qualtrics survey accessed through Amazon's Mechanical Turk worker interface. Participants were shown 30 summaries of research findings. Summaries were designed to break down the multi-cue non-generic used in Studies 3a-3c to further understand at what point participants' ratings of non-generic summaries drop off, that is, at what point they differ from bare generics. Five varieties of summaries were created (presented here with one example each, with emphases added):

- Bare: People with dysphoria **are** less sensitive to positive information in the environment.
- Simple past-tense non-generic: People with dysphoria **were** less sensitive to positive information in the environment.
- Qualifier (with past tense): People with dysphoria **were** less sensitive to positive information in the environment, **under certain circumstances**.
- Some (with past tense): **Some** people with dysphoria **were** less sensitive to positive information in the environment.
- Multi-cue non-generic: **Some** people with dysphoria **were** less sensitive to positive information in the environment, **under certain circumstances**.

As in Studies 3b and 3c, participants in Study 3d were asked to provide one rating (randomly assigned across participants) on a 1 to 7 scale per summary: "How important is this finding" (n = 151) or "How much would you want to draw conclusions from this finding?" (n = 148). Each participant received 6 summaries in each of the five forms. As in Studies 3a-3c, a given summary was presented in only one form per participant, but in different forms across participants.

Study 3d results. See Figure S2 and Table S7. A significant interaction between generic language and question was observed ($p < .001$); therefore separate models were run for each question. For the importance question, MTurk participants rated bare generics (M = 4.68, 95% CI = 4.53, 4.83) more highly than qualified non-generics (M = 4.54, 95% CI = 4.39, 4.70; b = -0.19, SE = 0.08, z = -2.24, p = .025) and multi-cue non-generics (M = 4.52, 95% CI = 4.36, 4.67; b = -0.24, SE = 0.08, z = -2.89, p = .004). No difference was observed for past-tense non-generics (M = 4.65, 95% CI = 4.49, 4.81; b = -0.03, SE = 0.08, z = -0.40, p = .692) or "some" non-generics (M = 4.63, 95% CI = 4.47, 4.78; b = -0.09, SE = 0.08, z = -1.05, p = .293).

For the conclude question, MTurk participants rated bare generics (M = 4.33, 95% CI = 4.16, 4.50) more highly than "some" non-generics (M = 4.06, 95% CI = 3.87, 4.24; b = -0.36, SE = 0.09, z = -4.29, $p < .001$), qualified non-generics (M = 3.95, 95% CI = 3.77, 4.12; b = -0.51, SE = 0.09, z = -5.93, $p < .001$), and multi-cue non-generics (M = 3.87, 95% CI = 3.67, 4.06; b = -0.59, SE = 0.09, z = -6.83, $p < .001$), and less highly than past-tense non-generics (M = 4.49, 95% CI = 4.31, 4.66; b = 0.18, SE = 0.09, z = 2.07, p = .039).

Studies 4a-4b supplemental method

Participants. See Table S8. Participants were recruited by Amazon's Mechanical Turk with the criterion that they were located in the United States, native English speakers, had not completed a previous version of this study, and had completed at least 100 HITs with a 95% approval rating (Study 4a: n = 202; Study 4b: n = 205). Participants were paid $0.50 for completing the study. These studies were approved under the same IRB protocol as Studies 2 and 3. Our target sample size for each study and test question was 100 participants. Studies were conducted in April and May 2019.

Based on the criteria established in Study 2, participants were excluded from the analysis if they completed fewer than 80% of the study items. 28 participants were excluded on this basis (4a = 0, 4b = 28). An additional 13 participants were excluded if they reported that they were not native English speakers (4a = 6, 4b = 7). Although we set U.S. location as a recruitment criterion, 1 participant in Study 3d who was a native English speaker reported that they were not currently in the U.S.; we included the data for that participant.

Study 4a procedure. Participants completed a Qualtrics survey accessed through Amazon's Mechanical Turk worker interface. Participants were told "Scientists write about their research findings in many different ways. Some ways of writing may make

the findings sound ["more important" or "sound more conclusive"] than other ways of writing." Participants were then were asked to directly compare a pair of summaries based on importance (n = 104) or how much they would want to conclude from the findings (n = 98; randomly assigned between subjects). Summaries had the same content and varied only in whether they were described using a bare generic (e.g., "People with dysphoria **are** less sensitive to positive information in the environment.") or a past-tense non-generic (e.g., "People with dysphoria **were** less sensitive to positive information in the environment."). From the set of 36 phrases used in Study 3, 12 pairs were randomly selected by Qualtrics for each participant. Summaries were labeled as "A" or "B". Within participants, summaries were ordered such that half of the pairs had A as the bare generic and half of the summaries had B as the bare generic. Participants compared the summaries on a 1 (A > B) to 7 (B > A) scale, with the midpoint (4) noted as the summaries being equal. Responses were scored such that higher scores indicted that bare generics were rated higher and past-tense non-generics were rated lower. To do this required reverse-coding trials on which A was the bare generic.

Study 4a results. See Table S9, top. To examine whether participants rated bare generics or past-tense non-generics more highly, we performed one-sample t-tests (chance = 4, the midpoint of the scale) for each question (importance, conclude). For participants who were asked to report how important the finding was, bare generics were rated more highly than would be expected by chance, M = 4.42, 95% CI = 4.22, 4.62, $t(103) = 4.07$, $p < .001$, $d = 0.40$. For participants who were asked how much they would conclude from the finding, ratings did not differ from chance, M = 4.07, 95% CI = 3.84, 4.31, $t(97) = 0.63$, $p = .53$, $d = 0.06$.

Study 4b procedure. Participants completed a Qualtrics survey accessed through Amazon's Mechanical Turk worker interface. Participants were given the same instructions as in Study 4a (importance: n = 102; conclude: n = 103). Instead of comparing 12 pairs of bare generic vs. past-tense non-generic summaries, participants compared 35 pairs of summaries in which bare generics were each contrasted with one of five forms: Framed generics (e.g., "**This study confirms that** people with dysphoria **are** less sensitive to positive information in the environment." Emphases added.), past-tense non-generics, qualifier non-generics, "some" non-generics, and multi-cue non-generics (7 of each version). Each participant received 7 exemplars with each of the five contrasts (i.e., bare generic vs. each of the five other forms). A given summary was presented in only one comparison per participant, but in different comparisons across participants. Otherwise, the procedure and coding were identical to that of Study 4a.

Study 4b results. See Figure S3 and Table S9, bottom. To examine whether participants rated bare generics or other alternatives more highly, we performed one-sample t-tests (chance = 4, the midpoint of the scale) for each question (importance, conclude). For participants who were asked to report how important the finding was, framed generics were rated higher than bare generics, M = 3.32, 95% CI = 3.07, 3.56, $t(101) = -5.61$, $p <$ .001, but bare generics were rated more highly than all other alternatives (ps < .007). For participants who were asked how much they would conclude from the finding, framed

generics were again rated higher than bare generics, $M = 3.22$, 95% $CI = 2.96, 3.49$, $t(102) = -5.85$, $p < .001$, as were non-generics with "some" or qualifiers added ($ps < .05$).

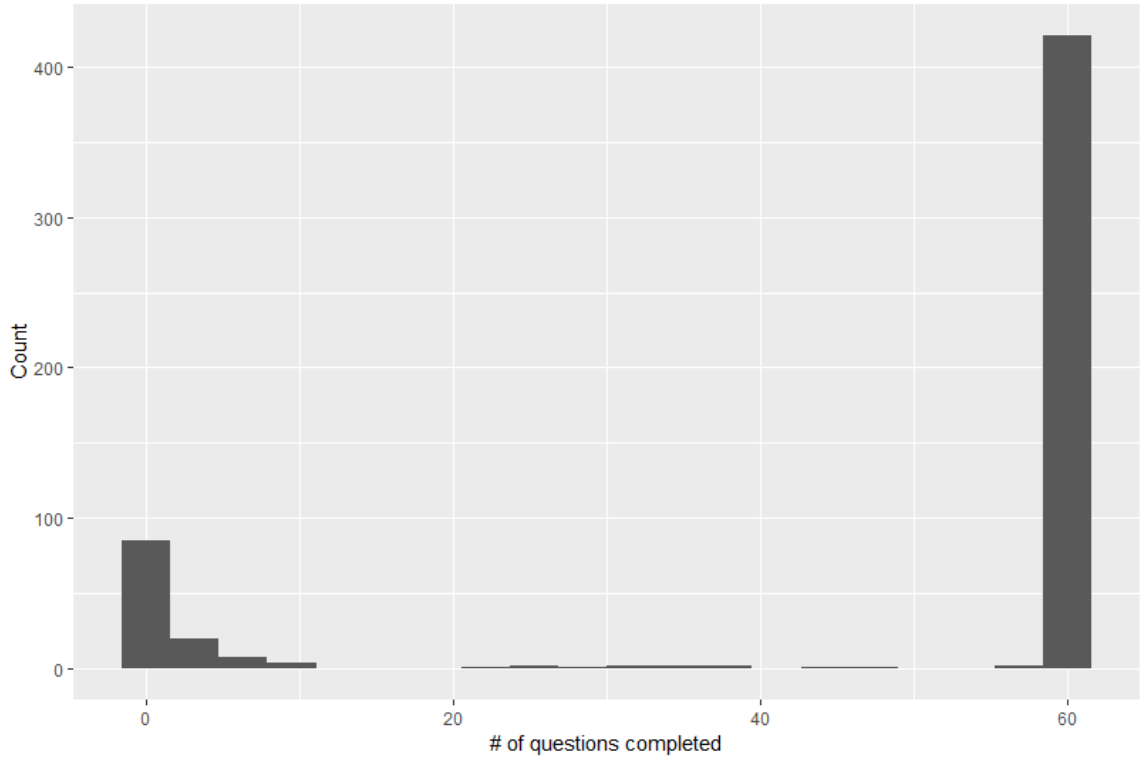

Fig. S1. Number of questions answered by each participant who submitted a survey in Study 2. Participants were excluded from analyses if they completed fewer than 48 of the 60 items.

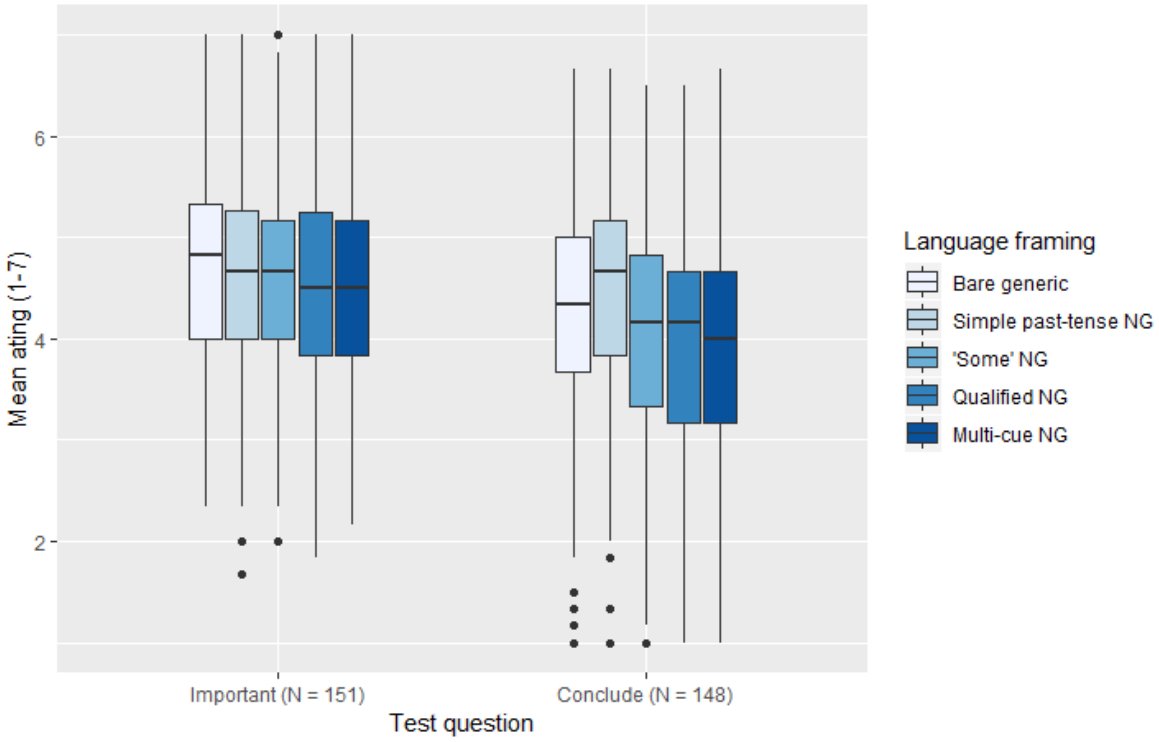

Fig S2. Study 3d, participants' ratings of the importance of research findings (left) or how much they would want to conclude from research findings (right), rated on a 1-7 scale (Study 3d). Participants were shown research summaries with subtle language variations, ranging from bare generics to multi-cue non-generics.

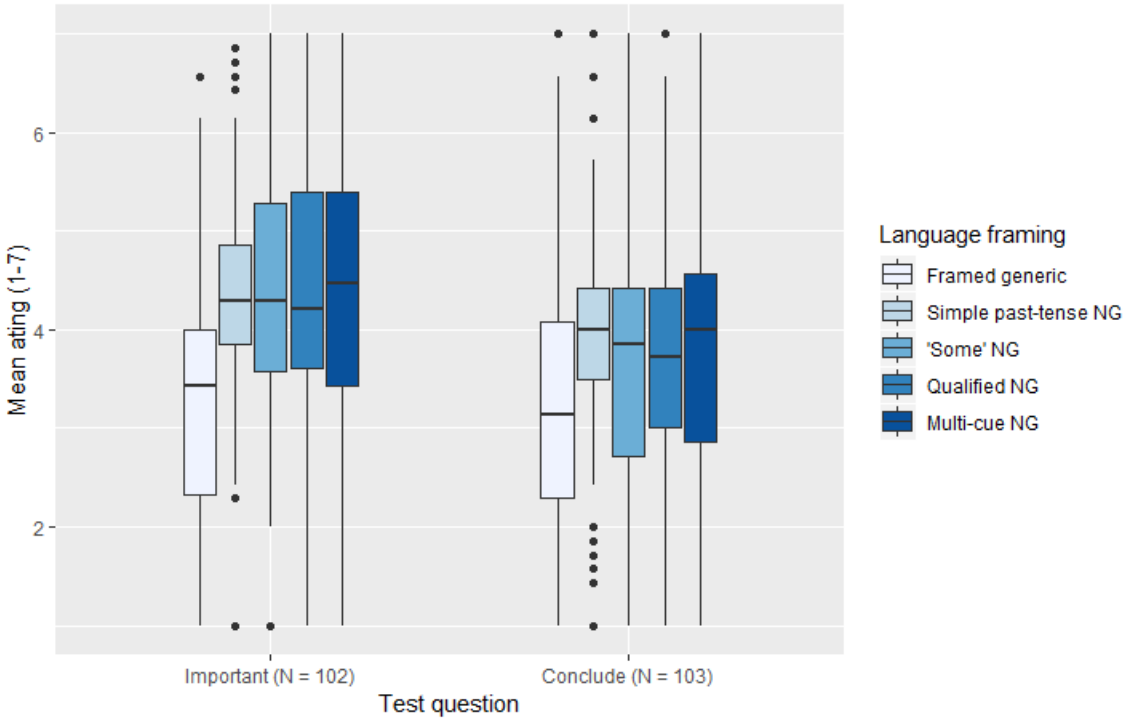

Fig S3. Study 4b, participants' direct comparisons of bare generics to an alternative (framed generics, simple past-tense NG, non-generics with "some", qualified non-generics, multi-cue non-generics, and simple past-tense non-generics,) in Study 4b.

Table S1. Study 1, Number of included and excluded articles from each journal.

Reason for exclusion	BP	Cog	CP	DS	IJP	J Ab Psy	JCCP	JECP	JESP	JML	PNAS	Total
Commentary, discussion, editorial	3	6	0	2	15	5	5	0	1	2	3	42
Correction	0	1	0	2	0	5	1	0	0	0	0	9
Animals only	0	5	0	3	0	0	0	0	0	0	7	15
Model, simulation	0	4	2	3	0	0	0	0	0	0	5	14
Reanalysis	0	0	0	0	4	1	0	0	2	1	0	8
Language corpus	0	1	1	0	0	0	0	0	0	0	2	4
No highlights	1	62	0	3	1	40	0	12	2	0	0	121
Total excluded	4	79	3	13	20	51	6	12	5	3	17	213
Total included	137	166	30	83	156	55	109	162	96	64	91	1149

Table S2. Journal requirements and descriptions from author guidelines available in August 2018 and general suggestions from the *Publication Manual of the American Psychological Association* (6th Edition).

Journal (Publisher)	Titles	Highlights/Summary Statements	Abstracts
Biological Psychology **(Elsevier)**	Concise and informative	A short collection of bullet points that convey the core findings of the article. 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).	A concise and factual abstract is required. The abstract should state briefly the purpose of the research, the principal results and major conclusions. An abstract is often presented separately from the article, so it must be able to stand alone. The abstract should be no more than 150 words.
Cognition **(Elsevier)**	Concise and informative	A short collection of bullet points that convey the core findings of the article. 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).	A concise and factual abstract is required. The abstract should state briefly the purpose of the research, the principal results and major conclusions. An abstract is often presented separately from the article, so it must be able to stand alone.
Cognitive Psychology **(Elsevier)**	Concise and informative	A short collection of bullet points that convey the core findings of the article. 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).	A concise and factual abstract is required. The abstract should state briefly the purpose of the research, the principal results and major conclusions. An abstract is often presented separately from the article, so it must be able to stand alone.
Developmental Science **(Wiley)**	A short informative title that contains the major key words	Up to four 'Research Highlights'; bulleted points outlining the key contributions to research the paper	Abstracts should be in the form of a continuous narrative, rather than divided into distinct sections (i.e., Background, Methods,

		makes. Each research highlight should not be longer than 25 words.	Results, Conclusions). No more than 250 words, containing the major keywords.
International Journal of Psychophysiology **(Elsevier)**	Concise and informative	A short collection of bullet points that convey the core findings of the article. 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).	An abstract is often presented separately from the article, so it must be able to stand alone. This should provide a concise description of the purpose of the report or review article and should not exceed 250 words.
Journal of Abnormal Psychology **(APA)**	No specific instructions	General Scientific Summaries: A brief (2-3 sentences) statement that, in nontechnical language, explains the contributions of the paper. Assume that the reader is an intelligent, interested individual who might know something about abnormal psychology, but may not know technical terms or abbreviations.	All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page.
Journal of Counseling and Clinical Psychology **(APA)**	The title of a manuscript should be accurate, fully explanatory, and preferably no longer than 12 words. The title should reflect the content and population studied (e.g., "treatment of	2–3 brief sentences regarding the public health significance of the study or meta-analysis described in their paper. This description should be included within the manuscript on the abstract/keywords page. It should be written in language that is easily understood by both professionals and members of the lay public.	Please include an Abstract of up to 250 words, presented in paragraph form. The Abstract should be typed on a separate page (page 2 of the manuscript), and must include each of the following sections: Objective: A brief statement of the purpose of the study Method: A detailed summary of the participants (N, age, gender, ethnicity) as well as descriptions of the study design, measures

	generalized anxiety disorders in adults").		(including names of measures), and procedures Results: A detailed summary of the primary findings that clearly articulate comparison groups (if relevant), and that indicate significance or confidence intervals for the main findings Conclusions: A description of the research and clinical implications of the findings
Journal of Experimental Child Psychology **(Elsevier)**	Concise and informative	A short collection of bullet points that convey the core findings of the article. 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).	A concise and factual abstract is required (Maximum words = 250). The abstract should state briefly the purpose of the research, the principal results and major conclusions.
Journal of Experimental Social Psychology **(Elsevier)**	Concise and informative	A short collection of bullet points that convey the core findings of the article. 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).	Abstracts should be no more than 250 words.
Journal of Memory & Language **(Elsevier)**	Concise and informative	A short collection of bullet points that convey the core findings of the article. 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point).	A concise and factual abstract is required of approximately 150 words. The abstract should state briefly the purpose of the research, the principal results and major conclusions. An abstract is often presented separately from the article, so it must be able to stand alone.

Proceedings of the National Academies of Science	Titles should be no more than three typeset lines (generally 135 characters including spaces) and should be comprehensible to a broad scientific audience. The specific organism studied should be included.	Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of specialty. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership.	Provide an abstract of no more than 250 words on page 2 of the manuscript. Abstracts should explain to the general reader the major contributions of the paper.
Publication Manual of the American Psychological Association **(6th Edition)**	No more than 12 words	No entry in the book's index	150-250 words for an abstract (check journal guidelines)

Table S3. Study 1, Article-level descriptives. For the number of participants, median and SE are provided. For all other variables, the N and % are provided.

	BP	Cog	CP	DS	IJP	J Ab Psy	JCCP	JECP	JESP	JML	PNAS	Total
Impact factor												
2015	3.23	3.41	4.54	3.98	2.60	5.54	4.71	2.33	2.50	5.22	9.42	n/a
2016	3.07	3.41	3.72	4.60	2.58	4.13	4.59	2.60	2.16	3.07	9.99	n/a
Median # of participants	40 (19.76)	82.5 (397.32)	139 (46.44)	73 (15.72)	36.5 (18.16)	340 (633.78)	183 (233.31)	97 (8.88)	367.5 (37.09)	123 (37.23)	101 (280.79)	93 (72.68)
Test country												
Unspecified	36 (26%)	47 (28%)	3 (10%)	23 (28%)	29 (19%)	16 (29%)	48 (44%)	37 (23%)	21 (22%)	16 (25%)	15 (16%)	291 (25%)
U.S. only	22 (16%)	42 (25%)	13 (43%)	23 (28%)	33 (21%)	23 (42%)	43 (39%)	46 (28%)	47 (49%)	28 (44%)	34 (37%)	354 (31%)
Not just U.S.	79 (58%)	77 (46%)	14 (47%)	37 (45%)	94 (60%)	16 (29%)	18 (17%)	79 (49%)	28 (29%)	20 (31%)	42 (46%)	504 (44%)
Participant Race												
Unspecified	111 (81%)	149 (90%)	29 (97%)	58 (70%)	139 (89%)	22 (40%)	37 (34%)	97 (60%)	63 (66%)	63 (98%)	70 (77%)	838 (73%)
Specified	26 (19%)	17 (10%)	1 (3%)	25 (30%)	17 (11%)	33 (60%)	72 (66%)	65 (40%)	33 (37%)	1 (2%)	21 (23%)	311 (27%)
Participant SES												
Unspecified	117 (85%)	155 (93%)	29 (97%)	59 (71%)	138 (88%)	35 (64%)	39 (36%)	94 (58%)	92 (96%)	63 (98%)	85 (93%)	906 (79%)
Specified	20 (15%)	11 (6%)	1 (3%)	24 (29%)	18 (12%)	20 (36%)	70 (64%)	68 (42%)	4 (4%)	1 (2%)	6 (7%)	243 (21%)
Participant Language												
Unspecified	114 (83%)	105 (63%)	22 (73%)	54 (65%)	143 (92%)	46 (84%)	81 (74%)	106 (65%)	85 (89%)	18 (28%)	79 (87%)	853 (74%)
Specified	23 (17%)	61 (37%)	8 (27%)	29 (35%)	13 (8%)	9 (16%)	28 (26%)	56 (35%)	11 (11%)	46 (72%)	12 (13%)	296 (26%)

Table S4. Study 2, Demographic characteristics for included participants (*N* = 416). For age and time on task, the median and SE are provided. For all other variables, the *N* and % are provided.

Participant characteristics	Study 2, *N* (%)
Question	
Importance	108 (26%)
Generalizability	108 (26%)
Sample size	103 (25%)
Diversity	97 (23%)
Total	416
Age (years): Median (SE)	38 (0.58)
Time on task (min): Median (SE)	12.68 (2.46)
Gender	
Women	227 (55%)
Men	189 (45%)
Other	0 (0%)
Race/ethnicity	
White	326 (78%)
Black	31 (7%)
Asian	30 (7%)
Hispanic/Latino	8 (2%)
Native American	1 (<1%)
More than one race/ethnicity	19 (5%)
Other	1 (<1%)
Education	
High school diploma/GED	58 (14%)
Some college	84 (20%)
Associates degree	60 (14%)
Bachelor's degree	161 (39%)
Some graduate school	11 (3%)
Master's degree	28 (7%)
Ph.D./M.D./J.D.	12 (3%)
Did not enter	2 (<1%)
Income	
Less than $10,000	23 (6%)
$10,000 - $19,999	31 (7%)
$20,000 - $29,999	42 (10%)
$30,000 - $39,999	61 (15%)
$40,000 - $49,999	52 (13%)
$50,000 - $59,999	52 (13%)
$60,000 - $69,999	40 (10%)
$70,000 - $79,999	28 (7%)
$80,000 - $89,999	20 (5%)
$90,000 - $99,999	15 (4%)
$100,000 - $149,999	37 (9%)

$150,000 or more	13 (3%)
Did not enter	2 (<1%)
Ever taken a psychology course	
Yes	232 (56%)
No	184 (44%)
Reads science-related materials	
None/did not enter	65 (16%)
Scientific journals	22 (5%)
Popular press science articles	153 (37%)
Science-related books	38 (9%)
Journals and popular press	15 (4%)
Popular press and books	42 (10%)
Journals and books	4 (1%)
Journals, popular press, books	75 (18%)

Table S5. Study 2, Regression tables for each test question. For importance, sample size, and diversity, data were analyzed using an ordinal regression model; for generalizability, data were analyzed using a linear regression model. Significant effects of generic language were observed for the importance, generalizability, and sample size questions and significant effects of content area were observed for all questions.

	Mean [95% CI]	Estimate	*SE*	*z value*	*p-value*
Importance					
Generic language: LR χ^2 (3) = 13.19, p = .004					
Bare generic	4.12 [3.97, 4.28]	0.15	0.06	2.34	.019
Framed generic	4.19 [4.04, 4.34]	0.20	0.06	3.19	.001
Hedged generic	4.18 [4.02, 4.34]	0.19	0.06	3.09	.002
Past-tense non-generic (reference)	4.03 [3.87, 4.18]	n/a	n/a	n/a	n/a
Content area: LR χ^2 (5) = 631.55, p < .001					
Biological	4.04 [3.87, 4.21]	-0.03	0.08	-0.39	.700
Clinical	5.09 [4.94, 5.23]	1.31	0.08	16.65	< .001
Cognitive	3.79 [3.61, 3.96]	-0.36	0.08	-4.67	< .001
Developmental	4.07 [3.89, 4.25]	-0.01	0.08	-0.14	.891
Social	3.73 [3.54, 3.92]	-0.44	0.08	-5.82	< .001
PNAS (reference)	4.06 [3.89, 4.23]	n/a	n/a	n/a	n/a
Generalizability					
Generic language					
Bare generic	54.58 [51.78, 57.38]	1.11	0.97	1.14	.254
Framed generic	55.70 [52.87, 58.53]	2.21	0.97	2.28	.023
Hedged generic	53.72 [50.91, 56.53]	0.24	0.97	0.25	.803
Past-tense non-generic (reference)	53.47 [50.55, 56.39]	n/a	n/a	n/a	n/a
Content area					
Biological	56.55 [53.63, 59.47]	1.27	1.19	1.07	.286
Clinical	52.45 [49.68, 55.22]	-2.86	1.19	-2.40	.017
Cognitive	53.13 [49.92, 56.35]	-2.17	1.19	-1.82	.069
Developmental	56.88 [53.66, 60.09]	1.57	1.19	1.32	.187
Social	51.88 [49.00, 54.76]	-3.43	1.19	-2.88	.004
PNAS (reference)	55.31 [52.14, 58.49]	n/a	n/a	n/a	n/a
Sample size					
Generic language: LR χ^2 (3) = 5.11, p = .164					
Bare generic	3.93 [3.76, 4.10]	0.11	0.07	1.66	.098
Framed generic	3.92 [3.74, 4.10]	0.11	0.06	1.71	.088
Hedged generic	3.94 [3.75, 4.12]	0.13	0.06	2.08	.038
Past-tense non-generic (reference)	3.86 [3.67, 4.04]	n/a	n/a	n/a	n/a
Content area: LR χ^2 (5) = 206.89, p < .001					
Biological	4.14 [3.94, 4.35]	0.32	0.08	4.03	< .001
Clinical	4.18 [3.98, 4.38]	0.35	0.08	4.41	< .001

Cognitive	3.61 [3.43, 3.79]	-0.43	0.08	-5.38	< .001
Developmental	3.59 [3.39, 3.79]	-0.45	0.08	-5.76	< .001
Social	4.04 [3.86, 4.22]	0.18	0.08	2.33	.020
PNAS (reference)	3.91 [3.73, 4.09]	n/a	n/a	n/a	n/a
Diversity					
Generic language: LR χ^2 (3) = 1.14, p = .768					
Bare generic	4.81 [4.59, 5.03]	0.03	0.07	0.41	.684
Framed generic	4.80 [4.59, 5.00]	-0.03	0.07	-0.42	.674
Hedged generic	4.79 [4.59, 4.99]	-0.04	0.07	-0.56	.579
Past-tense non-generic (reference)	4.80 [4.61, 5.00]	n/a	n/a	n/a	n/a
Content area: LR χ^2 (5) = 52.48, p < .001					
Biological	4.86 [4.63, 5.10]	-0.12	0.08	-1.39	.164
Clinical	4.76 [4.53, 4.98]	-0.28	0.08	-3.36	< .001
Cognitive	4.69 [4.47, 4.92]	-0.29	0.08	-3.48	< .001
Developmental	4.97 [4.75, 5.20]	-0.01	0.08	-0.08	.933
Social	4.55 [4.36, 4.75]	-0.48	0.08	-5.80	< .001
PNAS (reference)	4.95 [4.75, 5.14]	n/a	n/a	n/a	n/a

Table S6. Study 3, Demographic characteristics for included participants (*N* = 755). For age and time on task, the median and SE are provided. For all other variables, the *N* and % are provided.

Participant characteristics	Study 3a *N* (%)	Study 3b *N* (%)	Study 3c *N* (%)	Study 3d *N* (%)
Question				
Importance	n/a	135 (51%)	60 (51%)	151 (51%)
Conclude	n/a	129 (49%)	58 (49%)	148 (49%)
Total	74	264	118	299
Age (years): Median (SE)	37 (1.15)	33 (0.61)	19 (0.08)	31 (0.64)
Time on task (min): Median (SE)	15.75 (2.15)	7.19 (0.38)	20.58 (2.04)	6.88 (0.61)
Gender				
Women	29 (39%)	118 (45%)	66 (56%)	121 (40%)
Men	43 (58%)	132 (50%)	51 (43%)	176 (59%)
Other	0 (0%)	0 (0%)	0 (0%)	1 (<1%)
Did not report	2 (3%)	14 (5%)	1 (1%)	1 (<1%)
Race/ethnicity				
White	61 (82%)	192 (73%)	67 (57%)	196 (66%)
Black	2 (3%)	16 (6%)	11 (9%)	34 (11%)
Asian	3 (4%)	16 (6%)	27 (23%)	26 (9%)
Hispanic/Latino	1 (1%)	11 (4%)	1 (<1%)	20 (7%)
Native American	0 (0%)	5 (2%)	1 (<1%)	9 (3%)
More than one race/ethnicity	4 (5%)	11 (4%)	8 (7%)	13 (4%)
Other	0 (0%)	0 (0%)	2 (2%)	0 (0%)
Did not report	2 (3%)	14 (5%)	1 (<1%)	1 (<1%)
Education				
High school diploma/GED	7 (9%)	30 (11%)	0	36 (12%)
Some college	13 (18%)	64 (24%)	118	70 (23%)
Associates degree	14 (19%)	30 (11%)	0	39 (13%)
Bachelor's degree	28 (38%)	89 (34%)	0	124 (41%)
Some graduate school	0 (0%)	0 (0%)	0	0 (0%)
Master's degree	10 (14%)	37 (14%)	0	29 (10%)
Ph.D./M.D./J.D.	0 (0%)	0 (0%)	0	0 (0%)
Did not enter	2 (3%)	14 (5%)	0	1 (<1%)
Income				
Less than $10,000	3 (4%)	13 (5%)	0 (0%)	14 (5%)
$10,000 - $19,999	6 (8%)	20 (8%)	5 (4%)	13 (4%)
$20,000 - $29,999	13 (18%)	34 (13%)	2 (2%)	40 (13%)
$30,000 - $39,999	10 (14%)	30 (11%)	1 (1%)	40 (13%)
$40,000 - $49,999	6 (8%)	24 (9%)	4 (3%)	51 (17%)
$50,000 - $59,999	4 (5%)	25 (9%)	2 (2%)	31 (10%)
$60,000 - $69,999	4 (5%)	21 (8%)	7 (6%)	36 (12%)
$70,000 - $79,999	9 (12%)	18 (7%)	6 (5%)	25 (8%)
$80,000 - $89,999	4 (5%)	15 (6%)	4 (3%)	12 (4%)
$90,000 - $99,999	3 (4%)	8 (3%)	3 (3%)	13 (4%)

$100,000 - $149,999	8 (11%)	29 (11%)	28 (24%)	17 (6%)
$150,000 or more	2 (3%)	12 (5%)	52 (44%)	5 (2%)
Did not enter	2 (3%)	15 (6%)	4 (3%)	2 (1%)
Ever taken a psychology course				
Yes	35 (47%)	151 (57%)	118	167 (56%)
No	37 (50%)	98 (37%)		131 (44%)
Did not enter	2 (3%)	15 (6%)		1 (<1%)
Reads science-related materials				
None/did not enter	27 (36%)	70 (27%)	27 (23%)	47 (16%)
Scientific journals	1 (1%)	16 (6%)	4 (3%)	30 (10%)
Popular press science articles	24 (32%)	80 (30%)	32 (27%)	102 (34%)
Science-related books	6 (8%)	18 (7%)	3 (3%)	31 (10%)
Journals and popular press	3 (4%)	9 (3%)	13 (11%)	17 (6%)
Popular press and books	6 (8%)	29 (11%)	11 (9%)	21 (7%)
Journals and books	1 (1%)	1 (<1%)	4 (3%)	6 (2%)
Journals, popular press, books	6 (8%)	41 (16%)	24 (20%)	45 (15%)

Table S7. Study 3, Regression tables for each experiment (R package: ordinal, function: clmm).

	Mean	Estimate	*SE*	*z value*	*p-value*
Study 3a					
Generic x Question interaction: LR χ^2 (6) = 6.45, p = .375					
Generic language: LR χ^2 (2) = 405.41, p < .001					
Bare generic (reference)	4.52 [4.32, 4.72]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.54 [4.34, 4.74]	0.06	0.08	0.66	.511
Multi-cue non-generic	3.85 [3.59, 4.11]	-0.77	0.09	-8.92	< .001
Question: LR χ^2 (3) = 123.25, p < .001					
Importance	4.15 [3.95, 4.35]	-0.02	0.09	-0.21	.836
Generalize in U.S.	4.55 [4.30, 4.81]	0.51	0.08	6.06	< .001
Generalize outside U.S.	4.42 [4.15, 4.68]	0.37	0.08	4.44	< .001
Conclude (reference)	4.09 [3.89, 4.30]	n/a	n/a	n/a	n/a
Study 3b					
Generic x Question interaction: LR χ^2 (2) = 24.89, p < .001					
Importance: Generic language: LR χ^2 (2) = 101.21, p < .001					
Bare generic (reference)	4.53 [4.37, 4.68]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.56 [4.42, 4.69]	0.04	0.06	0.55	.585
Multi-cue non-generic	4.11 [3.94, 4.28]	-0.54	0.06	-8.40	< .001
Conclude: Generic language: LR χ^2 (2) = 233.27, p < .001					
Bare generic (reference)	4.38 [4.21, 4.55]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.42 [4.26, 4.58]	0.03	0.07	0.44	.661
Multi-cue non-generic	3.67 [3.48, 3.87]	-0.88	0.07	-13.01	< .001
Pairwise by question (C-I)		**Estimate**	***SE***	***z*-ratio**	***p*-value**
Bare generic	4.46 [4.34, 4.57]	-0.15	0.14	-1.06	.898
Past-tense non-generic	4.49 [4.39, 4.59]	-0.15	0.14	-1.09	.886
Multi-cue non-generic	3.90 [3.77, 4.03]	-0.55	0.14	-3.92	.001
Study 3c					
Generic x Question interaction: LR χ^2 (2) = 3.46, p = .177					
Generic language: LR χ^2 (2) = 177.88, p < .001					
Bare generic (reference)	4.60 [4.46, 4.74]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.59 [4.46, 4.72]	0.02	0.10	0.16	.874
Multi-cue non-generic	3.98 [3.81, 4.14]	-0.92	0.10	-8.88	< .001
Question: LR χ^2 (1) = 3.33, p = .068					
Importance	4.49 [4.32, 4.66]	0.25	0.18	1.37	.172
Conclude (reference)	4.27 [4.12, 4.43]	n/a	n/a	n/a	n/a
Study 3b vs. 3c					
Generic x Question interaction: LR χ^2 (2) = 26.80, p < .001					
Generic x Participants interaction: LR χ^2 (2) = 0.88, p = .645					
Importance:					
Generic language: LR χ^2 (2) = 180.17, p < .001					
Bare generic (reference)	4.58 [4.45, 4.70]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.59 [4.48, 4.70]	0.02	0.05	0.29	.776
Multi-cue non-generic	4.11 [3.98, 4.25]	-0.61	0.05	-11.44	< .001
Participants: LR χ^2 (1) = 0.58, p = .448					

MTurk	4.40 [4.26, 4.53]	-0.12	0.162	-0.76	.448
Student (reference)	4.49 [4.32, 4.66]	n/a	n/a	n/a	n/a
Conclude:					
Generic language: LR χ^2 (2) = 327.37, p < .001					
Bare generic (reference)	4.42 [4.29, 4.55]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.45 [4.33, 4.57]	0.02	0.05	0.44	.659
Multi-cue non-generic	3.71 [3.55, 3.87]	-0.87	0.06	-15.46	< .001
Participants: LR χ^2 (1) = 0.48, p = .487					
MTurk	4.16 [4.02, 4.30]	-0.11	0.15	-0.70	.486
Student (reference)	4.27 [4.12, 4.43]	n/a	n/a	n/a	n/a
Pairwise by question (C-I)		**Estimate**	***SE***	***z*-ratio**	***p*-value**
Bare generic	4.50 [4.41, 4.59]	-0.18	0.11	-1.60	.600
Past-tense non-generic	4.52 [4.44, 4.60]	-0.17	0.11	-1.51	.659
Multi-cue non-generic	3.92 [3.82, 4.03]	-0.52	0.11	-4.62	< .001
Study 3d					
Generic x Question interaction: LR χ^2 (4) = 42.52, p < .001					
Importance: Generic language: LR χ^2 (4) = 11.93, p = .018					
Bare generic (reference)	4.68 [4.53, 4.83]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.65 [4.49, 4.81]	-0.03	0.08	-0.40	.692
"Some" non-generic	4.63 [4.47, 4.78]	-0.09	0.08	-1.05	.293
Qualified non-generic	4.54 [4.39, 4.70]	-0.19	0.08	-2.24	.025
Multi-cue non-generic	4.52 [4.36, 4.67]	-0.24	0.08	-2.89	.004
Conclude: Generic language: LR χ^2 (4) = 118.80, p < .001					
Bare generic (reference)	4.33 [4.16, 4.50]	n/a	n/a	n/a	n/a
Past-tense non-generic	4.49 [4.31, 4.66]	0.18	0.09	2.07	.039
"Some" non-generic	4.06 [3.87, 4.24]	-0.36	0.09	-4.29	< .001
Qualified non-generic	3.95 [3.77, 4.12]	-0.51	0.09	-5.93	< .001
Multi-cue non-generic	3.87 [3.67, 4.06]	-0.59	0.09	-6.83	< .001
Pairwise by question (C-I)		**Estimate**	***SE***	***z*-ratio**	***p*-value**
Bare generic	4.51 [4.40, 4.62]	-0.51	0.17	-3.02	.076
Past-tense non-generic	4.57 [4.45, 4.69]	-0.28	0.17	-1.68	.808
"Some" non-generic	4.34 [4.22, 4.47]	-0.82	0.17	-4.89	< .001
Qualified non-generic	4.25 [4.13, 4.37]	-0.89	0.17	-5.28	< .001
Multi-cue non-generic	4.19 [4.06, 4.32]	-0.92	0.17	-5.48	< .001

Table S8. Study 4, Demographic characteristics for included participants (*N* = 407). For age and time on task, the median and SE are provided. For all other variables, the *N* and % are provided.

Participant characteristics	Study 4a *N* (%)	Study 4b *N* (%)
Question		
Importance	104 (51%)	102 (50%)
Conclude	98 (49%)	103 (50%)
Total	202	205
Age (years): Median (SE)	30 (0.70)	31 (0.76)
Time on task (min): Median (SE)	4.22 (3.88)	8.35 (0.65)
Gender		
Women	67 (33%)	83 (40%)
Men	132 (65%)	118 (58%)
Other	3 (1%)	4 (2%)
Did not report		
Race/ethnicity		
White	150 (74%)	140 (68%)
Black	16 (8%)	17 (8%)
Asian	9 (4%)	9 (4%)
Hispanic/Latino	9 (4%)	20 (10%)
Native American	4 (2%)	3 (1%)
More than one race/ethnicity	11 (5%)	11 (5%)
Other	0 (0%)	1 (<1%)
Did not enter	3 (1%)	4 (2%)
Education		
High school diploma/GED	21 (10%)	35 (17%)
Some college	46 (23%)	50 (24%)
Associates degree	36 (18%)	25 (12%)
Bachelor's degree	75 (37%)	75 (37%)
Some graduate school	3 (1%)	0 (0%)
Master's degree	15 (7%)	10 (5%)
Ph.D./M.D./J.D.	3 (1%)	6 (3%)
Did not enter	3 (1%)	4 (2%)
Income		
Less than $10,000	8 (4%)	9 (4%)
$10,000 - $19,999	16 (8%)	18 (9%)
$20,000 - $29,999	22 (11%)	26 (13%)
$30,000 - $39,999	25 (12%)	28 (14%)
$40,000 - $49,999	24 (12%)	29 (14%)
$50,000 - $59,999	32 (16%)	32 (16%)
$60,000 - $69,999	14 (7%)	11 (5%)
$70,000 - $79,999	19 (9%)	8 (4%)
$80,000 - $89,999	10 (5%)	7 (3%)
$90,000 - $99,999	10 (5%)	2 (1%)

$100,000 - $149,999	13 (6%)	22 (11%)
$150,000 or more	5 (2%)	9 (4%)
Did not enter	4 (2%)	4 (2%)
Ever taken a psychology course		
Yes	114 (56%)	105 (51%)
No	86 (43%)	95 (46%)
Did not report	2 (1%)	5 (2%)
Reads science-related materials		
None/did not enter	45 (22%)	46 (22%)
Scientific journals	21 (10%)	19 (9%)
Popular press science articles	48 (24%)	52 (25%)
Science-related books	22 (11%)	20 (10%)
Journals and popular press	14 (7%)	13 (6%)
Popular press and books	13 (6%)	15 (7%)
Journals and books	6 (3%)	6 (3%)
Journals, popular press, books	33 (16%)	34 (17%)

Table S9. Study 4, Comparisons between bare generics and other alternatives (framed generics, non-generics, qualified non-generics, generics using "some", and multi-cue non-generics). Means above 4.0 and positive *t*-values represent higher ratings for bare generics; means below 4.0 and negative *t*-values mean higher ratings for the alternative listed.

	Mean [95% *CI*]	***t*-value**	***p*-value**	**Cohen's *d***
Study 4a				
Importance	4.42 [4.22, 4.62]	4.07	< .001	0.40
Conclude	4.07 [3.84, 4.31]	0.63	.53	0.06
Study 4b				
Importance				
Framed generic	3.32 [3.07, 3.56]	-5.61	< .001	0.56
Past-tense non-generic	4.39 [4.19, 4.59]	3.86	< .001	0.38
"Some" non-generic	4.43 [4.17, 4.69]	3.29	.001	0.33
Qualifier non-generic	4.39 [4.12, 4.66]	2.83	.006	0.28
Multi-cue non-generic	4.41 [4.13, 4.69]	2.89	.005	0.29
Conclude				
Framed generic	3.22 [2.96, 3.49]	-5.85	< .001	0.58
Past-tense non-generic	3.99 [3.78, 4.19]	-0.14	.892	0.01
"Some" non-generic	3.71 [3.42, 4.00]	-2.01	.047	0.20
Qualifier non-generic	3.71 [3.46, 3.97]	-2.24	.027	0.22
Multi-cue non-generic	3.80 [3.51, 4.09]	-1.35	.181	0.13

Supplemental References

1. R Core Team (2016) *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria).

2. Gelman SA, Raman L (2003) Preschool children use linguistic form class and pragmatic cues to interpret generics. *Child Dev* 74(1):308–325.