Journal of Personality and Social Psychology

Are Women Really (Not) More Talkative Than Men? A Registered Report of Binary Gender Similarities/Differences in Daily Word Use

Colin A. Tidwell, Alexander F. Danvers, Valeria A. Pfeifer, Danielle B. Abel, Eva Alisic, Andrew Beer, Sabrina J. Bierstetel, Kathryn L. Bollich-Ziegler, Michelle Bruni, William R. Calabrese, Christine Chiarello, Burcu Demiray, Sona Dimidjian, Karen L. Fingerman, Maximilian Haas, Deanna M. Kaplan, Yijung K. Kim, Goran Knezevic, Ljiljana B. Lazarevic, Minxia Luo, Alessandra Macbeth, Joseph H. Manson, Jennifer S. Mascaro, Christina Metcalf, Kyle S. Minor, Suzanne Moseley, Angelina J. Polsinelli, Charles L. Raison, James K. Rilling, Megan L. Robbins, David Sbarra, Richard B. Slatcher, Jessie Sun, Mira Vasileva, Simine Vazire, and Matthias R. Mehl

Online First Publication, January 13, 2025. https://dx.doi.org/10.1037/pspp0000534

CITATION

Tidwell, C. A., Danvers, A. F., Pfeifer, V. A., Abel, D. B., Alisic, E., Beer, A., Bierstetel, S. J., Bollich-Ziegler, K. L., Bruni, M., Calabrese, W. R., Chiarello, C., Demiray, B., Dimidjian, S., Fingerman, K. L., Haas, M., Kaplan, D. M., Kim, Y. K., Knezevic, G., Lazarevic, L. B., Luo, M., Macbeth, A., Manson, J. H., Mascaro, J. S., Metcalf, C., Minor, K. S., Moseley, S., Polsinelli, A. J., Raison, C. L., Rilling, J. K., Robbins, M. L., Sbarra, D., Slatcher, R. B., Sun, J., Vasileva, M., Vazire, S., & Mehl, M. R. (2025). Are women really (not) more talkative than men? A registered report of binary gender similarities/differences in daily word use. *Journal of Personality and Social Psychology*. Advance online publication. https://dx.doi.org/10.1037/pspp0000534

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Personality and Social Psychology:
Personality Processes and Individual Differences

© 2025 American Psychological Association
ISSN: 0022-3514

https://doi.org/10.1037/pspp0000534

Are Women Really (Not) More Talkative Than Men? A Registered Report of Binary Gender Similarities/Differences in Daily Word Use

Colin A. Tidwell[1], Alexander F. Danvers[2], Valeria A. Pfeifer[1], Danielle B. Abel[3], Eva Alisic[4], Andrew Beer[5], Sabrina J. Bierstetel[6], Kathryn L. Bollich-Ziegler[7], Michelle Bruni[8], William R. Calabrese[9], Christine Chiarello[8], Burcu Demiray[10], Sona Dimidjian[11], Karen L. Fingerman[12], Maximilian Haas[13, 14], Deanna M. Kaplan[15, 16], Yijung K. Kim[12], Goran Knezevic[17], Ljiljana B. Lazarevic[17], Minxia Luo[10], Alessandra Macbeth[18], Joseph H. Manson[19], Jennifer S. Mascaro[15, 16], Christina Metcalf[20], Kyle S. Minor[3], Suzanne Moseley[1], Angelina J. Polsinelli[21], Charles L. Raison[22], James K. Rilling[15, 16], Megan L. Robbins[8], David Sbarra[1], Richard B. Slatcher[23], Jessie Sun[24], Mira Vasileva[4], Simine Vazire[25], and Matthias R. Mehl[1]

[1] Department of Psychology, The University of Arizona
[2] Sierra Tucson, Tucson, Arizona, United States
[3] School of Science, Indiana University Indianapolis
[4] Melbourne School of Population and Global Health, The University of Melbourne
[5] Department of Psychology, University of South Carolina Upstate
[6] Department of Psychology, Franciscan University of Steubenville
[7] Department of Psychology, Seattle University
[8] Department of Psychology, University of California, Riverside
[9] Department of Psychiatry and Behavioral Health, Renaissance School of Medicine at Stony Brook University
[10] Department of Psychology, University of Zurich
[11] Department of Psychology and Neuroscience, University of Colorado Boulder
[12] Department of Human Development and Family Sciences, The University of Texas at Austin
[13] Department of Psychology and Education Sciences, University of Geneva
[14] Faculty of Psychology, UniDistance Suisse
[15] Department of Family and Preventive Medicine, Emory University
[16] Department of Spiritual Health, School of Medicine, Woodruff Health Sciences Center, Emory University
[17] Department of Psychology, University of Belgrade
[18] Department of Psychology, Azusa Pacific University
[19] Department of Anthropology, University of California, Los Angeles
[20] Department of Psychiatry, University of Colorado Anschutz Medical Campus
[21] Department of Neurology, Indiana University School of Medicine
[22] Department of Psychiatry, University Wisconsin-Madison
[23] Department of Psychology, University of Georgia
[24] Department of Psychological and Brain Sciences, Washington University in St. Louis
[25] School of Psychological Sciences, The University of Melbourne

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Richard E. Lucas served as action editor.

Colin A. Tidwell https://orcid.org/0000-0003-1748-2440

Colin A. Tidwell, Alexander F. Danvers, and Valeria A. Pfeifer contributed equally to this article. All authors other than Colin A. Tidwell, Alexander F. Danvers, Valeria A. Pfeifer, and Matthias R. Mehl are listed in alphabetical order. The raw data for reproducing the reported results are publicly available on the Open Science Framework at https://osf.io/wrtcz/. The audio recordings and the verbatim transcripts from which the word count variable is derived cannot be made available for reasons of protecting participants' privacy. The full analyses scripts for reproducing the reported results are publicly available on the Open Science Framework at https://osf.io/wrtcz/.

This project (including the original data collections) was funded by the National Institutes of Health and Centers for Disease Control and Prevention (Grants R01AG046460, P30AG066614, P2CHD042849, R01HD069498, 3R01AT004698, 5R01AT004698, R03 CA137975, BCS-1125553, and 100019_165572), the American Psychological Foundation Pearson Early Career Grant, John Templeton Foundation Positive Neuroscience Award, Mind and Life 1440 award, and the Ministry of Education, Science and Technological Development, Serbia (Grant 179018).

Colin A. Tidwell played a lead role in project administration and an equal role in conceptualization, data curation, methodology, writing–original draft, and writing–review and editing. Alexander F. Danvers played an equal role in conceptualization, methodology, formal analysis, and writing–original draft. Valeria A. Pfeifer played a lead role in visualization and validation and an equal role in formal analysis and writing–review and editing. Danielle B. Abel played a supporting role in data curation. Eva Alisic played a supporting role in data curation. Andrew Beer played a supporting role in data curation. Sabrina J. Bierstetel played a supporting role in data curation. Kathryn L. Bollich-Ziegler played a supporting role in data curation. Michelle Bruni played a supporting role in data curation. William R. Calabrese played a supporting role in data curation. Christine Chiarello played a supporting role in data curation. Burcu Demiray played a supporting role in data curation. Sona Dimidjian played a supporting role in data curation. Karen L. Fingerman played a supporting role in data curation. Maximilian Haas played a supporting role in data curation. Deanna M. Kaplan played a supporting role in data curation and writing–review and editing. Yijung K. Kim played a supporting role in data curation. Goran Knezevic played a supporting role in

continued

Women are widely assumed to be more talkative than men. Challenging this assumption, Mehl et al. (2007) provided empirical evidence that men and women do not differ significantly in their daily word use, speaking about 16,000 words per day (WPD) each. However, concerns were raised that their sample was too small to yield generalizable estimates and too age and context homogeneous to permit inferences beyond college students. This registered report replicated and extended the previous study of binary gender differences in daily word use to address these concerns. Across 2,197 participants (more than five-fold the original sample size), pooled over 22 samples (631,030 ambient audio recordings), men spoke on average 11,950 WPD and women 13,349 WPD, with very large individual differences (<100 to >120,000 WPD). The estimated gender difference (1,073 WPD; $d = 0.13$; 95% CrI [316, 1,824]) was about twice as large as in the original study. Smaller differences emerged among adolescent (513 WPD), emerging adult (841 WPD), and older adult (−788 WPD) participants, but a substantially larger difference emerged for participants in early and middle adulthood (3,275 WPD; $d = 0.32$). Despite the considerable sample size(s), all estimates carried large statistical uncertainty and, except for the gender difference in early and middle adulthood, provide inconclusive evidence regarding whether the two genders ultimately speak a practically equivalent number of WPD, based on the preregistered ±1,000 WPD regions of practical equivalence criterion. Experienced stress had no meaningful effect on the gender difference, and no clear pattern emerged as to whether the gender difference is accentuated for subjectively rated compared with objectively observed talkativeness.

Keywords: gender stereotypes, sex differences, lexical budget, daily vocabulary, replication

"The tongue is the sword of a woman and she never lets it become rusty" (Chinese proverb). "Women's tongues are like lambs' tails—they are never still" (English saying). "The North Sea will sooner be found wanting in water than a woman at a loss for words" (Danish saying. "A man a word, a woman a dictionary" (German saying). These, and similar popular sayings, suggest that a widespread and culturally deeply engrained stereotype exists that women are more talkative than men (especially when thinking of gender as binary). Scientifically, the existence (and persistence) of the stereotype has been confirmed in both qualitative (Talbot, 2003) and quantitative (Donovan, 2011) research. With respect to direct empirical evidence, one particularly relevant study asked participants to rate the degree to which they agreed with a list of adjectives representing common societal stereotypes of women on a 1–9-point Likert scale. Participants rated "talkative" as the trait they agreed with most highly (6.5) for all traits about women aside from "dependent" (Landrine, 1985).

The stereotype also gained widespread scientific and public attention in the first edition of neuropsychiatrist Louann Brizendine's book *The Female Brain* (2007). In the book, Brizendine wrote: "A woman uses about 20,000 words per day while a man uses about 7,000." Although not supported by empirical evidence, these numbers have since circulated widely throughout television, radio, and print media. Historically, the notion of daily lexical budgets was introduced 15 years prior, in the context of marriage counseling, as a way of illustrating gendered relationship dynamics (Liberman, 2006). Since then, it has become a pervasive fixture in arguments of gender differences in talkativeness. The pejorative nature of this stereotype makes evaluating its accuracy particularly important (Czopp et al., 2015; Schmader et al., 2008).

The first empirical data on the number of words men and women use on a daily basis were published by Mehl et al. in *Science* in 2007, a year after the publication of *The Female Brain*. In their study, Mehl et al. addressed a central measurement challenge of estimating how many words people use in a day by employing a novel ecological behavioral observation method, the electronically activated recorder (EAR). The EAR is a portable audio recorder that intermittently (e.g., five times per hour) records short (e.g., 30-s) ambient sound bites (Mehl et al., 2001). Participants wear the EAR while going about their day, unaware of when exactly it is recording. Through its person-centered tracking of ambient sounds, the EAR yields acoustic logs of participants' days and provides objective records of their activities, including their conversations. Through its sampling strategy, the EAR employs a representative design (i.e., samples situations representatively; Brunswik, 1955) and enables the study of larger numbers of participants (Schönbrodt & Perugini, 2013). The captured ambient sounds are then transcribed, and participants' daily word use is estimated from the number of recorded words.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

data curation and funding acquisition. Ljiljana B. Lazarevic played a supporting role in data curation and funding acquisition. Minxia Luo played a supporting role in data curation. Alessandra Macbeth played a supporting role in data curation. Joseph H. Manson played a supporting role in data curation. Jennifer S. Mascaro played a supporting role in data curation. Christina Metcalf played a supporting role in data curation. Kyle S. Minor played a supporting role in data curation and funding acquisition. Suzanne Moseley played a supporting role in data curation. Angelina J. Polsinelli played a supporting role in data curation and funding acquisition. Charles L. Raison played a supporting role in data curation. James K. Rilling played a supporting role in data curation. Megan L. Robbins played a supporting role in data curation. David Sbarra played a supporting role in data curation and writing–review and editing. Richard B. Slatcher played a supporting role in data curation. Jessie Sun played a supporting role in data curation and writing–review and editing. Mira Vasileva played a supporting role in data curation. Simine Vazire played a supporting role in data curation and writing–review and editing. Matthias R. Mehl played a lead role supervision and resources and an equal role in conceptualization, methodology, administration, funding acquisition, writing–original draft, and writing–review and editing.

Sona Dimidjian reported being a co-founder of Mindful Noggin, Inc. and receiving revenue from Mindful Noggin, Inc. Charles L. Raison worked as a consultant for Usona Institute, Otsuka, Novartis.

Correspondence concerning this article should be addressed to Matthias R. Mehl, Department of Psychology, The University of Arizona, 1503 East University Boulevard (Building 68), Tucson, AZ 85721, United States. Email: mehl@arizona.edu

Interestingly, and to the surprise of many, the analyses in the 2007 article revealed a gender difference of only 546 words per day (WPD), with women speaking an average of 16,215 words ($SD = 7{,}301$) and men an average of 15,669 WPD ($SD = 8{,}633$). This gender difference accounted for only 0.1% of the standardized variability (Cohen's $d = 0.07$, $r = .035$, $R^2 = .001$). Based on the study's sample size ($N = 396$), this effect size was far from statistically significant, ($p = .248$, one-tailed). The authors concluded that women and men effectively do not differ (much) in the number of words they utter on a daily basis and that, "on the basis of available empirical evidence, … the widespread and highly publicized stereotype about female talkativeness is unfounded" (Mehl et al., 2007, p. 82).

The study garnered substantial national and international media attention and was well received by both scientists and the general public. Nevertheless, more than a dozen years after its publication, it seems to have had little effect on weakening the perception that women are excessively verbose in everyday life. Evidence that the stereotype is "alive and well" abounds on the internet (e.g., in pertinent memes such as "haha get it, cause women talk a lot"; Reddit, 2018) and also surfaces regularly in the spotlight of public life (Kobayashi & Murakami, 2021; Mangalindan, 2017; McCurry, 2021). Revisiting the original study by Mehl et al. (2007) is also important from a scientific perspective as it has been subject to important critiques. First, while a sample size of $N = 396$ is large for a naturalistic observation project, it is ultimately too small if the goal is to provide strong evidence for the absence or presence of a gender difference in daily word use (Schönbrodt & Wagenmakers, 2018). Second, although one of their six analyzed samples was collected in Mexico, the majority of the data (87%) originated in one single city: Austin, Texas. This raises legitimate questions about the generalizability of the obtained estimates. Third, their sample consisted entirely of college students. Arguably, if the goal is to rule out biological, brain-based sex differences in talkativeness, as were postulated in *The Female Brain* ("All of this is hardwired into the brains of women. These are the talents women are born with that many men, frankly, are not"; Brizendine, 2007, p. 8), college students should be an adequate population. Nevertheless, it is without a doubt a critical limitation for the generalizability of the estimates. Fourth and last, an informal reanalysis of the data, published in *Psychology Today*, found that when a unique sample that was collected in the context of the 9/11 terrorist attacks is excluded from the analyses, the results show that women talk slightly more than men ($d = 0.13$; Schmitt, 2016). This suggests that it might be important to consider participants' levels of experienced stress, as biobehavioral coping processes can alter people's sociability and can do so in gender-linked ways (Taylor et al., 2000).

In addition to these critiques, new pertinent data have emerged since the study's publication in 2007. In the same year, Leaper and Ayres (2007) published a meta-analysis on gender differences in language use that included "talkativeness" as an outcome Across these 70 studies and 4,385 participants, men were more talkative than women ($d = -0.14$). However, when parsing the data by how talkativeness was operationalized, no effect ($d = 0.01$) emerged for the 13 studies that used a word count measure. The authors stated that the "studies in the meta-analysis were based mostly on formal tests of language ability rather than observations of actual conversations" (Leaper & Ayres, 2007, p. 329). A gender-linked aspect of language ability was also investigated by Schultheiss et al. (2021). Based on a very large sample (11,528 participants), they found meta-analytic evidence for a female advantage in narrative writing fluency. Women consistently wrote longer stories than men in a narrative writing task ($d = 0.31$), and this effect was mediated by the sex-dimorphic hormone estradiol, suggesting a potential biological basis. Finally, Onnela et al. (2014) used sociometric badges (which derive speech information from spectral ambient audio features) to estimate the talkativeness of men and women in the workplace. They found no overall gender difference, but women talked more in collaborative settings, and men talked more in noncollaborative settings. Taken together, the critiques voiced in response to the original study, and the inconclusive new data that have since emerged point to the importance of revisiting the original study: (a) to replicate it in a much larger and more diverse sample (to increase the statistical precision and generalizability of the estimates) and (b) expand on it by exploring the role of participant age (as a marker of developmental processes) and experienced stress (as a marker of biobehavioral coping processes).

A fruitful line of research investigates gender differences in talkativeness as they manifest in specific, theoretically defined conversation contexts. For example, in the 2014 book *The Silent Sex: Gender, Deliberation, and Institutions*, Karpowitz and Mendelberg found that "the ratio of female-to-male talk was largest when majority-rule groups contained a supermajority of women." In addition, as another example, a recent study examining gender differences in leadership emergence found that men tended to participate more in group conversations than women, suggesting that, in agentic communication contexts, talking can mark dominance (Badura et al., 2018). While context naturally matters in that it can shape how and how much individuals, and these two binary genders, talk in different situations, our research, as a replication of Mehl et al.'s (2007) study, focuses on perceived gender-related general talkativeness in relation to actual talking behavior across the natural range of daily contexts. Our approach addresses the stereotype at the general, context-encompassing level at which it socioculturally exists and follows Brunswik's (1956) representative design (i.e., the representative sampling of contexts from underlying ecologies, in this case a day in the life) to accomplish this. This way, the current project expands upon the existing literature by conducting a representative analysis of how many words humans in general, and men and women in particular, use in a day. This project therefore also serves to complement systematic, theoretical analyses of contextualized gender differences in talkativeness (Leaper & Ayres, 2007).

In addition, the (context-representative) number of words humans speak per day (and the variability therein) that Mehl et al.'s (2007) study yielded, and that this study seeks to update, is also of interest to other scientific fields. This is evidenced by the diverse citations to the original study (e.g., from linguistics, communication, cognitive science, evolutionary biology). In sum, context can play an important role in shaping talking behavior; at the same time, this study's approach of estimating the number of words spoken per day in relation to gender (through representative sampling across the range of daily contexts) is valuable for both theoretical (i.e., addressing the stereotype at the level at which it exists) and methodological (i.e., naturalistic observation of talkativeness) reasons.

Finally, it is important to recognize that our approach is unable to speak to the processes that may underlie a possible gender difference

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

in daily word use. For example, our approach cannot help identify to what extent a possible gender difference in daily word use may be due to biological versus sociological processes and to what extent it may result from the two genders proactively (or "inherently") selecting themselves into different daily contexts versus being reactively pulled or passively constrained into different contexts (e.g., due to societal norms or pressures). Systematic experimental approaches (which test specific theoretical hypotheses) or large-scale research syntheses (e.g., Leaper & Ayres, 2007) may ultimately be in a better position to accomplish this.

The primary goal of the present study was to conduct a registered replication of Mehl et al.'s (2007) study, estimating the gender difference in men's and women's daily word use. For this purpose, the first and last authors invited the principal investigators of existing EAR studies to join this replication project. After initially reaching out directly to selected (i.e., known to us) EAR researchers, resulting in the first 18 samples, a systematic search for additional published and unpublished studies yielded an additional four samples. These additional samples originated from emails sent to listservs of four professional societies (Society for Personality and Social Psychology, Association for Research in Personality, Society for Ambulatory Assessment, and Society for the Science of Clinical Psychology), increasing the overall sample size by 306 participants (16%). Our analyses relied on raw data from these studies (word count per day), which makes the present analysis a "mega-analysis" (Sung et al., 2014).

We only excluded studies that (a) used the EAR method but did not transcribe the captured conversations (i.e., relied exclusively on behavior coding); (b) did not complete data collection and processing (i.e., transcription) by March 1, 2022 (the time we pooled the data); and (c) did not obtain participant consent for analyzing the data beyond the original study aims. For this replication, we also excluded any data that were included in the original study. This way, we were able to obtain, harmonize, and pool data for 2,197 participants (five times more than the original sample size) originating from 22 different samples in four countries (the United States, Switzerland, Serbia, Australia) from individuals ranging in age from 10 to 94 years old.

In all studies, participants wore the EAR for multiple days, with a sampling of ambient audio occurring from morning to night. The sampling schedule (weekday and/or weekend), duration (number of days), frequency (recordings per hour), recording length (duration of each recording), and blackout period (i.e., nightly nonrecording) varied by study, as did the study aims (ranging from social, personality, clinical, health, developmental, and evolutionary psychology to anthropology and neuroscience). The EAR deployment, though, was highly similar across all studies, including the safety measures for protecting the privacy of participants and their conversation partners and ensuring the confidentiality of the data (Manson & Robbins, 2017; Mehl, 2017; Mehl & Holleran, 2007).

Research Question 1 (RQ1)

Is there a gender difference in words spoken per day between men and women? RQ1 is the direct registered replication of the estimates of male and female daily word use by Mehl et al. (2007). Addressing the study's main critiques that the sample was too small to yield precise and generalizable estimates, and too homogeneous with respect to age and context to permit inferences beyond college students, we seek to provide an updated estimate using our full sample (2,197 participants). Replicating Mehl et al.'s (2007) study, we expect to find no gender difference in how many words men and women speak per day (Hypothesis 1).

Research Question 2 (RQ2)

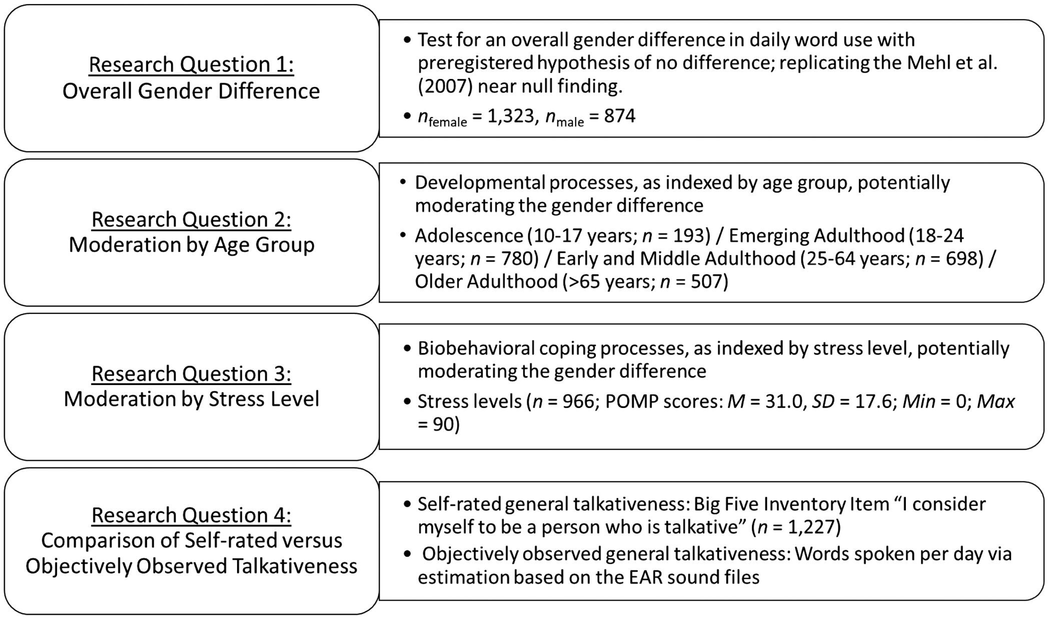
To what extent does age (as a marker of developmental processes) moderate a gender difference in words spoken per day between men and women? One of the main critiques of the original study was that its sample consisted entirely of college students and, thus, overwhelmingly of young adults. Theoretically, developmental processes may affect gender differences in talkativeness (Eagly et al., 2000; Taylor et al., 2000). Developmental processes can do so through biological mechanisms (e.g., sex hormones during puberty differentially affecting social brain maturation), social mechanisms (e.g., the college environment maximally affording opportunities or creating peer pressure to socialize, potentially "disguising" an underlying gender difference), or the interaction between the two (e.g., mobility and/or cognitive changes creating barriers to socialize among some older adults). Because our pooled data include participants from 10 to 94 years of age (see Figure 1 for age distribution), we can empirically evaluate evidence for how developmental processes potentially influence gender differences in daily word use. Although our design does not allow for a clean separation of age from developmental processes, the results can help constrain the range of plausible explanations.

Research Question 3 (RQ3)

To what extent does experienced stress (as a marker of biobehavioral coping processes) moderate a gender difference in words spoken per day between men and women? Another theoretical possibility is that gender differences in talkativeness might be modest in ordinary daily life but become accentuated in times of stress. Although Schmitt's reanalysis of Mehl et al.'s (2007) data found that the four male participants in the (very small; $N = 11$) September 11, 2001, sample talked more during this national upheaval than the seven female participants (Schmitt, 2016), a reverse pattern is more consistent with prior theorizing around the role of gender in responding to stress. Taylor et al.'s (2000) tend-and-befriend model implied that women and men might differ in their reactions to stress, as women's biological stress response may prime them toward affiliation and increased speech (as compared with a fight-or-flight response that would include less speech). Supporting research has found cross-sectional associations between different stress and tending-and-befriending measures in adult women, such as between cardiovascular stress (i.e., blood pressure) and partner touching and oxytocin levels (Light et al., 2005) and between hormonal (i.e., cortisol) and relationship stress and oxytocin levels (Taylor et al., 2006). Also consistent with the idea that women might socialize to mitigate stress, another meta-analysis found that women used verbal expressions to others as a coping strategy (to seek emotional support) more so than men (Tamres et al., 2002).

Note that tend-and-befriend processes can (and likely do) also unfold at the within-person level. Tend-and-befriend theory, though,

Figure 1
Sample Structure and Research Questions



Note. The data are pooled for 2,197 participants from 22 samples; in all studies, participants wore the EAR for multiple days, and it intermittently recorded ambient sound bites from their daily lives. EAR = electronically activated recorder; POMP = percent of maximum possible; Min = Minimum; Max = Maximum.

is first and foremost concerned with systematic between-person, specifically between-gender, variability in affiliation under stress. Our research question follows this logic and therefore proposes to test experienced stress as a plausible moderator of the gender difference in daily word use.

Across the studies, participants naturally experienced a wide range of stress levels around the time of wearing the EAR. Some studies monitored participants specifically in normatively stressful times (e.g., after a recent divorce, during adjuvant breast cancer treatment, postpartum, after a child's injury). Other studies monitored them during presumed "normal" times. In both cases, some study participants experienced high and others modest or low levels of stress around the time their conversations were being sampled. Because stress measures were available for n = 966 participants (44% of the full sample), we can empirically evaluate evidence for how biobehavioral coping processes potentially influence the gender difference in daily word use.

Research Question 4 (RQ4)

How do gender differences compare for objectively observed versus subjectively rated general talkativeness? Finally, a unique psychometric opportunity emerged in this project from the fact that several studies included a personality assessment using the Big Five Inventory (John et al., 1991). The Big Five Inventory includes an item that asks participants to provide self-reports of how talkative they are ("I see myself as someone who is talkative"; *strongly disagree* to *strongly agree*). This subjective measure of self-rated general talkativeness complements the main objective measure of observed talkativeness. It is conceivable that talkativeness looks different from the inside than from the outside (Vazire, 2010; Vazire & Mehl, 2008). Importantly, in this regard, David Schmitt's analyses on the Psychology Today website (March 17, 2016) found, using data from this item, that women describe themselves as more talkative than men (d = 0.27). For Research Question 4, we will estimate the gender difference in self-rated general talkativeness in the full sample as well as in all the subanalyses for RQ2 and RQ3. Figure 1 shows a schematic overview of the sample and the research questions.

Two methodological questions in this context concern (a) whether these research questions are more appropriately addressed meta-analytically or via a secondary data analysis and (b) whether registration is appropriate given the research team's prior access to the data. Because our research questions specifically afford analyses at the person level for the independent variable (i.e., participant-level gender rather than sample-level gender composition information), proposed potential moderators (i.e., participant-level age and stress level rather than sample-level age and stress summary statistics), and control variable (amount of EAR data available per participant), and because we were able to obtain access to the raw data, we opted in favor of a secondary analysis of pooled, raw, participant-level data (also known as "mega-analysis," e.g., Sung et al., 2014). Further, registering our secondary analyses helps guard against potential (confirmation) bias toward replicating our prior finding of no substantial gender differences (e.g., via the implicit use of

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

researchers' degrees of freedom). Our adopted approach is in line with best practices for preregistration of secondary data analysis (e.g., van den Akker et al., 2021; Weston et al., 2019).

Method

This is a registered replication report. The version of the article that received in-principle acceptance, along with the corresponding preregistration (including the analysis plan), is publicly available on the Open Science Framework at https://osf.io/d6t53. All data, analysis code, and supplemental materials are publicly available on the Open Science Framework at https://osf.io/wrtcz/.

Ethics Information

The individual protocols for each of the included 22 samples were approved by the respective principal investigators' institutional review boards. All analyses were conducted on deidentified data collected from participants who consented to having their data used in future studies and for aims other than the ones of the study in which they participated.

Design

All samples included in this study employed ambulatory assessment designs that used the EAR as a naturalistic observation method. For the Stage 1 registered report, we assembled the full pooled data set. To calibrate our research questions against the available data (e.g., to ensure adequate sample size per group), we reviewed univariate descriptive statistics for all variables except those comprising our target outcome variable: words per day. Importantly, we did not compute the outcome variable, words per day, from its constituting elements before we had received in-principle acceptance and preregistered the project (https://osf.io/d6t53).

Sampling Plan

Prior to any exclusions, this study comprises a sample size of 2,323 participants. These participants come from 22 samples spanning 14 years of data collection (2005–2019) across four countries (the United States, Switzerland, Serbia, Australia; Table 1). We excluded participants before conducting any analyses pertaining to the research questions (i.e., examining only the distributions of individual variables). We excluded a total of 126 participants. Eighty participants were excluded because of missing EAR data (defined here as no word count and/or no valid waking files), 37 were excluded due to mental health diagnoses (i.e., schizophrenia) with criteria impacting speech production and processing, six were excluded because they did not report their gender, and an additional three were excluded because their self-reported gender did not fall along the gender binary, which is necessary to replicate the analyses from the original study. The full sample size after these exclusions is N = 2,197 and 631,030 recordings. The effective sample size for the analyses depends on the availability of other demographic information (e.g., age, stress level; see Figure 1).

Samples

Sample 1

As part of a larger study on daily experiences and well-being strategies, 303 older adult participants wore the EAR for 5–6 days. The EAR recorded for 30 s every 7 min during waking hours. Data collection occurred in the greater Austin, Texas, metropolitan statistical area between 2016 and 2017 (Fingerman et al., 2020).

Sample 2

As part of a larger study on personality and interpersonal roles, 299 college students wore the EAR for 6–8 days. The EAR recorded for 30 s every 9.5 min between the hours of 7 a.m. and 2 a.m. Data collection occurred in St. Louis, Missouri, between 2012 and 2013 (Sun & Vazire, 2019).

Sample 3

As part of a larger study on the effects of two meditation interventions on daily behavior, 182 adult participants wore the EAR twice for 3 days each (separated by 4 weeks). The EAR recorded for either 50 s every 9 min or 30 s every 12 min during waking hours. Data collection occurred in Atlanta, Georgia, and Tucson, Arizona, between 2010 and 2013 (Kaplan et al., 2022).

Sample 4

As part of a larger study on real-world cognitive activities and conversational time travel, 109 young and older adults wore the EAR for 2 weekdays and 2 weekend days. The EAR recorded at random times for 30 s, on average every 12 min, during an 18-hr daytime period. Data collection occurred in Zurich, Switzerland, between 2014 and 2015 (Luo et al., 2021).

Sample 5

As part of a larger study on personality and behavior, 108 college students wore the EAR for 2 weekdays. The EAR recorded for 30 s every 12 min between the hours of 7:30 a.m. and 11:30 p.m. Data collection occurred at the University of South Carolina, Upstate between 2009 and 2010 (Beer & Vazire, 2017).

Sample 6

As part of a larger study on social and cognitive behavior and aging, 107 older adults wore the EAR for 2 weekdays and 2 weekend days. The EAR recorded for 30 s every 12 min during waking hours. Data collection occurred in Tucson, Arizona, between 2015 and 2017 (Polsinelli et al., 2020).

Sample 7

As part of a larger study on daily behavior and life history strategy, 89 college students wore the EAR for 3 days. The EAR recorded randomly for 30 s every 12 min between 6 a.m. and 12 a.m. Data collection occurred at the University of California, Los Angeles between 2013 and 2015 (Manson, 2018).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1
Overview of the Samples Included in the Analyses

Sample	*N*	% Women	M_{age}	% White	Age range	Stress level POMP score *M* (*SD*)	Number of days of EAR monitoring	Location of data collection	Years of data collection	Participant demographics	Reference publication
1	303	53.1	74.1	70.3	65–92		5–6	Austin, Texas	2016–2017	Older adults in the community	Fingerman et al. (2020)
2	299	68.6	19.2	52.2	18–29	27.0 (13.7)	6–8	St. Louis, Missouri	2012–2013	Undergraduate students	Sun and Vazire (2019)
3	182	66.5	33.6	53.3	25–55	35.4 (11.8)	3	Atlanta, Georgia and Tucson, Arizona	2010–2013	Adults in the community	Kaplan et al. (2022)
4	109	57.8	44		18–83		4	Zurich, Switzerland	2014–2015	Young and older adults	Luo et al. (2021)
5	108	75.0	22.4	53.9	18–54		2	Columbia, South Carolina	2009–2010	Undergraduate students	Beer and Vazire (2017)
6	107	54.2	75.8	99.1	65–90		4	Tucson, Arizona	2015–2017	Older adults in the community	Polsinelli et al. (2020)
7	89	55.1	20.1	16.9	19–40		3	Los Angeles, California	2013–2015	Undergraduate students	Manson (2018)
8	69	84.1	20.1	100	19–28		3	Belgrade, Serbia	2015–2018	Undergraduate students	Lazarević et al. (2020)
9	64	60.9	20.3	71.9	18–36		2	Indianapolis, Indiana	2014–2016	Adults with and without schizotypy	Minor et al. (2018)
10	55	0.0	33.1	69.1	22–46		2	Atlanta, Georgia	2011–2013	New fathers	Mascaro et al. (2018)
11	47	53.3	35	73.3	24–51	43.5 (15.5)	3	Austin, Texas	2006–2007	Adult couples	Bierstetel and Slatcher (2020)
12	150	42.0	12.9	23.3	10–18	13.5 (12.0)	4	Metro Detroit, Michigan	2010–2014	Children with asthma	Farrell et al. (2018)
13	120	71.7	44.0	63.9	21–65	42.1 (18.0)	9	Tucson, Arizona	2011–2015	Adults experiencing a divorce	O'Hara et al. (2020)
14	36	50.0	44.1	61.8	20–64		2	Indianapolis, Indiana	2015–2019	Adults without schizophrenia	Abel et al. (2021)
15	52	59.6	57.2	84.6	24–94	33.0 (14.9)	3	Tucson, Arizona	2007–2011	Women with breast cancer	Robbins et al. (2014)
16	45	100.0	29.9	91.1	22–39		3	Boulder, Colorado	2014–2015	Postpartum women	Metcalf and Dimidjian (2020)
17	43	46.5	12.8		10–16	43.7 (22.2)	2	Melbourne, Australia	2013–2014	Children recovering from an injury	Alisic et al. (2017)
18	13	100.0	55.6	92.3	40–83	24.0 (22.2)	3	Tucson, Arizona	2005–2006	Adults with rheumatoid arthritis	Robbins et al. (2011)
19	73	65.8	44.73	82.2	20–79		4	Buffalo, New York	2013–2014	Adults in the community	Calabrese et al. (2024)
20	81	76.5	44.5	84.0	19–81	43.5 (19.5)	3	Geneva, Switzerland	2018–2019	Young and older adults	Haas et al. (2022)
21	75	66.7	19.20	10.7	18–25		4	Riverside, California	2017–2019	Undergraduate students	Macbeth et al. (2022)
22	77	58.4	32.16	41.9	18–66	27.0 (16.4)	4	Southern California	2014–2018	Adult couples	Robbins et al. (2024)

Note. The sample sizes reflect the participants whose data were analyzed for this project (so, the postexclusion sample sizes). POMP = percent of maximum possible; EAR = electronically activated recorder.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Sample 8

As part of two studies concerned with the ambulatory assessment of language use, 69 undergraduate students wore the EAR for 3 days. The EAR recorded for 30 s every 6 min between 9 a.m. and 12 a.m. Data collection occurred in Belgrade, Serbia, between 2015 and 2018 (Lazarević et al., 2020).

Sample 9

As part of a larger study on social behavior and schizotypy, 64 undergraduate students with low and high schizotypy wore the EAR for 2 days. The EAR recorded for 5 min 12 times per day between 6 a.m. and 12 a.m. Data collection occurred in Indianapolis, Indiana, between 2014 and 2016 (Minor et al., 2018).

Sample 10

As part of a larger study on the biological bases of paternal nurturance, 55 fathers wore the EAR for 2 days. The EAR recorded for 50 s every 9 min between 8 a.m. on a Sunday and 8 a.m. on a Tuesday (to record 1 workday and 1 nonworkday). Data collection occurred in Atlanta, Georgia, between 2011 and 2013 (Mascaro et al., 2018).

Sample 11

As part of a larger set of studies on interpersonal conflict and diurnal cortisol patterns, 47 adults wore the EAR for 3 days. The EAR recorded for 120 s every 12 min (but only the first 50 s of every recording were transcribed and coded by research assistants). Data collection occurred in Austin, Texas, between 2006 and 2007 (Bierstetel & Slatcher, 2020; Slatcher & Robles, 2012).

Sample 12

As part of a larger study on Asthma in the Lives of Families Today, 150 youth and their caregivers wore the EAR for 4 days (2 weekdays and 2 weekend days). The EAR recorded for 50 s every 9 min during waking hours. Only data from the youths in the sample were included in our study to ensure independence between parents' and their children's EAR files. Youths' files were selected to improve the sample size for this group. Data collection occurred in the Metro Detroit region of the United States between 2010 and 2014 (Farrell et al., 2018).

Sample 13

As part of a larger study on divorce, sleep, and daily social environment, 120 adult participants wore the EAR three times for 3 days (Friday to Sunday), separated by 2 months each. The EAR recorded for 30 s every 12 min during waking hours. Data collection occurred in Tucson, Arizona, between 2011 and 2015 (O'Hara et al., 2020).

Sample 14

As part of a larger study to understand real-world social functioning deficits in schizophrenia, 36 control participants (without schizophrenia) wore the EAR for 2 days. The EAR recorded for 5 min every 90 min between 6 a.m. and 12 a.m. Thirty-seven participants with a schizophrenia diagnosis were excluded from the analyses because of the potential impact that this condition (and its medical treatment) can have on speech production and processing. Data collection occurred in Indianapolis, Indiana, between 2015 and 2019 (Abel et al., 2021).

Sample 15

As part of a study on the daily life of couples coping with breast cancer, 52 breast cancer patients and their cohabitating partners wore the EAR for 3 days (Friday to Sunday). Within each couple, one member was randomly chosen to avoid statistical nonindependence. The final sample consisted of 27 breast cancer patients and 25 partners. The EAR recorded for 50 s every 9 min during the couples' waking hours. Data collection occurred in Tucson, Arizona, between 2007 and 2011 (Robbins et al., 2014).

Sample 16

As part of a larger study on social–emotional aspects of daily life in postpartum women, 49 participants wore the EAR for 3 days (Friday to Sunday). Four participants were excluded in accordance with our exclusion criteria. The EAR recorded for 30 s every 12.5 min between 6 a.m. and 12 a.m. Data collection occurred in Boulder, Colorado, between 2014 and 2015 (Metcalf & Dimidjian, 2020).

Sample 17

As part of a larger study on the daily life of children following an injury, 43 children and adolescents wore the EAR for 2 days when the child was mainly at home (such as a weekend or holiday). The EAR recorded for 30 s every 5 min during waking hours. Data collection occurred in Melbourne, Australia, between 2013 and 2014 (Alisic et al., 2015).

Sample 18

As part of a larger study on coping with rheumatoid arthritis in daily life, 13 adults wore the EAR twice for 3 days (Friday to Sunday), 1 month apart. The EAR recorded for 50 s every 18 min during waking hours. Data collection occurred in Tucson, Arizona, between 2005 and 2006 (Robbins et al., 2011).

Sample 19

As part of a study on the measurement of personality disorder patterns and psychosocial dysfunction, 73 adults wore the EAR for 4 consecutive days between a Thursday at 5 p.m. and a Tuesday at 2 a.m. The EAR recorded for 30 s every 12.5 min. The data were collected via the Computerized Adaptive Test for Personality Disorder Study at the University at Buffalo, New York, between 2013 and 2014 (Calabrese et al., 2024).

Sample 20

As part of a study examining the age-prospective memory paradox via novel real-world assessment technologies, a total of 81 participants, 43 younger adults (aged 19–32) and 38 older adults (aged 60–81), wore the EAR for 3 days. The EAR recorded for 30 s every 12 min on average between the hours of 7 a.m. and 9 p.m. Data collection occurred at the University of Geneva in Switzerland, between 2018 and 2019 (Haas et al., 2022).

Sample 21

As part of a larger study on the day-to-day linguistic experiences of young adults, 75 undergraduate participants who spoke a variety of languages (including, but not limited to, English, Vietnamese, and Spanish) wore the EAR for 4 days, which included 2 weekdays and 2 weekend days. All transcripts were translated into English to estimate the daily word count consistent with the other samples included in this study. The EAR recorded for 40 s every 12 min. Data collection occurred at the University of California, Riverside, between the years of 2017 and 2019 (Macbeth et al., 2022).

Sample 22

As part of a larger study on similarities and differences in social interaction quality and social network size, 154 participants in same- and different-gender couples wore the EAR for two weekends, separated by 1 month. Within each couple, one member was randomly chosen to ensure statistical nonindependence (however, prioritizing participants who completed both study time points in couples where one member was missing one). The final sample consisted of 77 participants. The EAR recorded for 50 s every 9 min and 25 s on average. Data collection occurred throughout Southern California between the years of 2014 and 2018 (Robbins et al., 2024).

Measures

Gender

Gender was analyzed binarily as either man (coded as 0) or woman (coded as 1).

Daily Word Use

The number of words that participants spoke per day was estimated following the protocol established by Mehl et al. (2007). For this, only EAR sound files in which participants were deemed awake and wearing the EAR were used ("valid waking files"). For these files, participants' speech (and only their speech) was transcribed by human transcribers, and the verbatim transcripts were text analyzed using the Linguistic Inquiry and Word Count software (Pennebaker et al., 2015) to count the number of words that each participant uttered. The number of words spoken per day was estimated by (a) calculating the average number of words that a participant spoke per EAR recording (based on their number of valid waking files) and (b) extrapolating to the number of words spoken per day (using the study's recording length and an estimate of waking hours). For example, if a participant had 3,200 words recorded over the course of the study, across 400 valid waking recordings, the participant spoke eight words per EAR recording. With a recording length of 30 s, this would be estimated to, on average, 960 words per hour and, assuming 17 hr of time awake, 16,320 WPD.

Note that participants' actual waking hours cannot be determined directly from the EAR recordings because of differences in the studies' daily monitoring start and end times and nightly EAR recording blackout periods. Therefore, the number of words spoken per day is calculated using an epidemiological estimate of daily waking hours as multiplier of the number of words spoken per hour, which is calculated directly and empirically for each participant from their average number of words sampled per recording period (e.g., 30 s). This procedure followed the procedure employed in the original study. Also following the original study procedures, and further supported by a recent consensus statement by the American Academy of Sleep Medicine and Sleep Research Society (Watson, Badr, Belenky, Bliwise, Buxton, Buysse, Dinges, Gangwisch, Grandner, Kushida, Malhotra, Martin, Patel, Quan, & Tasali, 2015; Watson, Badr, Belenky, Bliwise, Buxton, Buysse, Dinges, Gangwisch, Grandner, Kushida, Malhotra, Martin, Patel, Quan, Tasali, et al., 2015), 17 hr was used as an estimate of daily waking hours for all participants 18 years or older (based on the lower bound of 7 hr recommended sleep for this age group; n = 1,985). Following the complementary consensus statement by the American Academy of Sleep Medicine for pediatric populations (Paruthi et al., 2016), 16 hr was used as an estimate of daily waking hours for participants 10–17 years of age (based on the lower bound of 8 hr recommended sleep for this age group; n = 193).

Amount of Available EAR Data

As control variables that were used for sensitivity analyses, we computed the amount of audio data that were available for each participant. The amount of audio data was available for estimating the daily word use dependent on the studies' sampling parameters including the duration of one recording (e.g., 30 s, 40 s, or 50 s or 5 min), the sampling frequency (e.g., every 6, 12, or 18 min), and the length of the monitoring (e.g., 2, 3, or 6 days) as well as the participants' sleep behavior and compliance. The available number of minutes of ambient sound recordings was computed by multiplying the obtained number of valid (i.e., compliant), waking (i.e., not-sleeping) sound files by the duration of one recording (in minutes). On average, participants had a little less than 3 hr of net recordings (M = 164.2 min, SD = 81.6 min).

Because the total recording time does not consider the time period over which the ambient audio recordings were gathered (e.g., 100 min of recording obtained within 2 days is presumably less representative than 100 min of recording spread over 5 days), we further estimated the net hours of EAR monitoring for each participant. We calculated this variable from the obtained number of valid, waking sound files, and the programmed number of recordings per hour (e.g., five times per hour if the EAR recorded every 12 min). On average, participants underwent 46.4 hr of net EAR monitoring (SD = 21.6). The net hours of EAR monitoring were highly correlated with the total net recording time, r = .78, 95% CI [0.76, 0.79].

Self-Reported Talkativeness

Information on participants' self-reported general talkativeness is taken from the first item of the 44-item Big Five Inventory ("I see myself as someone who is talkative"; John et al., 1991). This information was available in Samples 1, 2, 3, 4, 5, 13, 15, and 18 (n = 1,227). To harmonize this measure across forms of administration

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

in the different studies (e.g., 5- vs. 7-point scale), we converted all raw scores into percent of maximum possible (POMP) scores (P. Cohen et al., 1999).

Experienced Stress

Stress level information was available in Samples 2, 3, 11, 12, 13, 15, 17, 18, 20, and 22 ($n = 966$). Specifically, the Perceived Stress Scale (S. Cohen et al., 1983) was available for participants in samples 3, 11, 13, 15, 18, 20, and 22. The total number of acute stressors from the Youth Life Stress Interview (Krackow & Rudolph, 2008) was available for participants in Sample 12. The Child Revised Impact of Events Scale–13 (Perrin et al., 2005) was available for participants in Sample 13. Sample 2 used experience sampling (ESM) to measure perceived stress by including a single-item measure of how stressful participants' momentary situation was on a 1–5-point Likert scale. Participants completed the ESM protocol for 2 weeks but only wore the EAR the first week. To closely match the stress and talkativeness data, only ESM reports from the days in between the start and end of the EAR sampling period were included. All sampled ESM reports were then averaged into an overall measure of currently experienced stress (Sun & Vazire, 2019).

Based on theoretical considerations around the tend-and-befriend model (i.e., more stress-induced socializing for women), measures of current/recent stress were chosen in studies where other measures (e.g., early or cumulative life stress) were available. To harmonize the scores across the different scales and studies, the raw stress scores were again converted into POMP scores.

Self-Reported EAR Obtrusiveness and Compliance

Participants completed a standard eight-item self-report questionnaire on their experiences with the EAR (c.f., Mehl & Holleran, 2007). On a 5-point scale ranging from 1 (*not at all*) to 5 (*a great deal*), they rated the obtrusiveness of the EAR for themselves (e.g., "To what degree were you generally aware of the EAR?" "To what degree did the EAR impede on your daily activities?") and people around them (e.g., "To what degree were people around you aware of the EAR?" "To what degree did the EAR influence the behavior of people around you?"). Finally, they estimated the percent of their time awake when they were not wearing the EAR. The questionnaire was available in Samples 2, 4, 5, 7, 9, 13, 14, 15, 16, 18, 19, 21, and 22 ($n = 1{,}126$ participants; 51.3% of the sample) and can be found at https://osf.io/2tx35. The data are available at https://osf.io/wrtcz/.

Analysis Plan

All analyses were conducted at the level of the individual participant to maximally use the information contained in the data (e.g., age group and stress level). The analysis plan is summarized in Table 2.

RQ1

Because our study aimed to provide evidence regarding the presence or absence of a gender difference, and because our data had a nested structure (participants nested within samples), we used Bayesian multilevel modeling analyses. Specifically, we used Bayesian multilevel assessment of null values via regions of practical equivalence (ROPE; Kruschke, 2011, 2018).

With respect to specifying the limits of a ROPE, Kruschke (2018) argued:

> Because the ROPE is a decision threshold that captures practical equivalence, its limits are influenced by practical considerations. … Any decision rule must be calibrated to be useful to the audience of the analysis and to the people who are affected by the decision. (Kruschke, 2018, p. 276)

In scientific practice, effect size-based approaches to specifying the ROPE are common; researchers often use $\delta \pm .10$ based on the rule of thumb that one can think of "no effect" as less than half the size of a small effect:

> Cohen suggested that 0.2 is a "small" effect, and therefore we might say that an effect is practically equivalent to zero if it is less than, say, half the size of a small effect and falls within a ROPE of ±0.1 (Kruschke, 2018, p. 276)

On the other hand, effect size-based approaches are ultimately a "fallback convention when there is no way to calibrate effects" (Kruschke, 2018, p. 276).

One feature of the EAR method at the measurement level is that, through the representative sampling and behavior counting approach, it yields variables with nonarbitrary and intuitive metrics, in this case, the estimated number of words a person speaks in a day (Mehl, 2017). Nonarbitrary and inherently meaningful (based on personal experience) metrics facilitate the interpretation of effect sizes and calibration of psychological effects (Blanton & Jaccard, 2006; Sechrest et al., 1996). Therefore, a viable option here—and the option chosen—is to use the original, unstandardized metric, rather than a metric based on the standardized difference between the means to determine what one might consider a trivial gender difference in words spoken per day (Mehl et al., 2007).

Determining the maximum daily word use difference that should be considered practically equivalent is, of course, to some extent subjective. Considering different scenarios, we settled on a ±1,000-word ROPE because (a) it aligns well with the original effect size estimate from Mehl et al.'s (2007) report (women spoke about 546 WPD more than men), (b) it aligns well with an effect size-based approach to determining the ROPE (extrapolating from the original study data, a $\delta \pm .10$ difference should translate to roughly ±800 words), and (c) the general public tends to construe the magnitude of the gender difference in daily word use in multiples of one thousand words (e.g., 20,000 vs. 7,000 words), suggesting that anything less than 1,000 words would likely be considered trivial (e.g., 15,900 vs. 15,100 words). We also believe that 1,000 words is a conservative threshold given the numbers that have circulated in the media (cf. Liberman, 2006). Finally, we believe that self-replications of an original null result should select a realistic but "tight" threshold. For example, a 2,000-word difference (e.g., 17,000 vs. 15,000 words) might not be particularly meaningful. However, it broadens the ROPE for determining practical equivalence biases toward successful replication. Having to commit to (and justify) a definitive ROPE prior to the analyses is key in which the registered report format guards against confirmation bias through post hoc (implicit) use of researchers' degrees of freedom.

Gender difference estimates for which the 95% high-density interval (HDI) fell completely within a ±1,000-word ROPE centered

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Analysis Plan for Addressing the Research Questions

Question	Hypothesis	Sampling plan	Analysis plan	Interpretation given to different outcomes
RQ1: Is there a gender difference in daily word use between men and women?	We expected to find no gender difference in how many words men and women speak per day.	Bayesian assessment of null values via ROPE analysis and Cohen's *d* estimates	• 95% HDI using a ±1,000-word ROPE • Cohen's *d* of small ($d \leq 0.20$), medium ($d \leq 0.50$), and large ($d \leq 0.80$)	For gender differences, the difference coefficient is being tested: • 95% HDI falls completely within a ±1,000-word ROPE centered around a zero difference: Practically equivalent. • 95% HDI falls completely outside of a ±1,000-word ROPE: Support for gender difference. • 95% HDI falls partially within and partially outside a ±1,000-word ROPE: Inconclusive evidence.
RQ2: To what extent is age (as a marker of developmental processes) associated with the gender difference in daily word use between men and women?	Exploratory, no specific hypothesis preregistered	Bayesian assessment of null values via ROPE analysis and Cohen's *d* estimates	• 95% HDI using a ±1,000-word ROPE • Cohen's *d* of small ($d \leq 0.20$), medium ($d \leq 0.50$), and large ($d \leq 0.80$)	For gender differences by age group, each gender difference by age group coefficient is being tested: • See above (RQ1) for interpretations of the 95% HDIs.
RQ3: To what extent is experienced stress (as a marker of biobehavioral coping processes) associated with the gender difference in daily word use between men and women?	Exploratory, no specific hypothesis preregistered	Bayesian multilevel regression analysis and Cohen's *d* estimates	• 95% HDI of the Stress × Gender interaction • Cohen's *d* of small ($d \leq 0.20$), medium ($d \leq 0.50$), and large ($d \leq 0.80$)	For gender differences by stress level, the gender difference by stress level interaction is being tested: • 95% HDI of the interaction includes zero: No credible effect of stress. • 95% HDI of the interaction excludes zero: Direction and magnitude of the effect as indicated by the effect size estimate (Cohen's *d*).
RQ4: How do the gender difference effects compare for objectively observed versus subjectively rated general talkativeness?	Exploratory, no specific hypothesis preregistered	Bayesian multilevel regression analysis and Cohen's *d* estimate	• 95% HDI of the gender effect • Cohen's *d* of small ($d \leq 0.20$), medium ($d \leq 0.50$), and large ($d \leq 0.80$) • Descriptive effect size comparison	For differences with self-report, each of the previous coefficients is being tested (as above) with self-reported talkativeness as the outcome; the effect is interpreted using the 95% HDI; the effect sizes for subjectively rated talkativeness is being compared with the effect sizes obtained for objectively observed talkativeness.

Note. RQ = research question; ROPE = region of practical equivalence; HDI = high-density interval.

around a zero difference were interpreted as practically equivalent; those for which the 95% HDI fell completely outside of a ±1,000-word ROPE were interpreted as support for the existence of a gender difference; and those for which the 95% HDI fell partially within and partially outside a ±1,000-word ROPE were interpreted as providing inconclusive evidence. If the analysis yielded support in favor of a gender difference, the effect size was interpreted using Cohen's guidelines for a small ($d \leq 0.20$), medium ($d \leq 0.50$), and large ($d \leq 0.80$) effects (Zell et al., 2015).

RQ2

To capture how developmental processes might be associated with gender differences in talkativeness, we binned the sample into four subgroups reflecting the following four (roughly) consensually recognized developmental stages: (a) adolescence (10–17 years; n = 193), (b) emerging adulthood (18–24 years; n = 780), (c) early and middle adulthood (25–64 years; n = 698), and (d) older adulthood (>65 years; n = 507). This binning follows recommended age boundaries for the developmental stages and ensures that each bin has a large enough subsample size to yield sound estimates. We decided in favor of age binning relative to analyzing age continuously because it appears to better capture the "soft discontinuity" of developmental processes. For RQ2, we therefore estimated the gender difference separately for the four age groups. We then followed the procedure outlined for RQ1 to determine (a) whether a meaningful gender difference existed in each group (using the ±1,000-word ROPE) and, (b) if so, what the magnitude of the estimated effect was (using Cohen's guidelines). RQ2 went beyond the registered replication of the original study and was exploratory in nature. Because of the lack of strong prior evidence, we registered no specific predictions.

RQ3

For RQ3, we tested the extent to which the gender difference was moderated by participants' stress levels. We again followed the procedures for RQ1 to determine (a) whether a meaningful gender difference exists as moderated by participant stress level (using the ±1,000-word ROPE) and, (b) if so, what the magnitude of the estimated effect was (using Cohen's guidelines). RQ3 went beyond the registered replication of the original study and was exploratory in nature. Because of the lack of strong prior evidence, we registered no specific predictions.

RQ4

To compare effects for self-rated and objectively observed talkativeness, all the analyses performed above were repeated on the "I consider myself to be a talkative person" item from the Big Five Inventory (in samples that contain that item, n = 1,227). This involved estimating the gender difference for self-reported talkativeness overall, as moderated by age group, and as moderated by stress level. The same analysis strategies described above were employed (with the only difference that the raw metric was a difference in POMP scores, accompanied by a Cohen's d). To create an estimate of the difference between self-rated and objectively observed talkativeness, the two variables could be standardized and entered as a common outcome in a model with a random intercept term to account for the nesting of variables within participants. However, this would have added an additional interaction term for each test, turning one-way effects into two-way interactions and two-way interactions into three-way interactions. These types of higher order estimates notoriously require much larger sample sizes to obtain reliable estimates. We therefore compared the effect sizes obtained for self-rated (POMP score difference) and objectively observed talkativeness (words-per-day difference) descriptively by evaluating their respective magnitudes (using Cohen's standard guidelines for effect sizes). RQ4 went beyond the registered replication of the original study and was exploratory in nature. Because of the lack of strong prior evidence, we registered no specific predictions.

Sensitivity and Robustness Testing

Although the 22 samples compiled here all originated within studies that employed the EAR method, their underlying procedures differed in aspects that could potentially influence the results. These include the sampling frequency (e.g., every 5 min vs. 12 min vs. 90 min), the length of one recording (30 sec vs. 50 sec. vs. 5 min), the number of days over which data were collected (e.g., 2 days vs. 5 days vs. 7 days), and the proportion of sampled days that were weekend days (e.g., 2 weekdays and 2 weekend days: 0.5). These factors vary at Level 2, the sample level. In addition, the available audio data, that is, the number of minutes of recording (M = 164.2 min, SD = 81.6 min), and the number of hours over which the EAR monitoring occurred (M = 46.4 hr, SD = 21.6), are important methodological factors. These two variables vary at Level 1, the participant level.

We decided to use the following three variables for the sensitivity analyses (see Table 3 for deviations from the preregistration): (a) the *total recording time* (the net, i.e., awake and compliant, number of minutes of recording that the EAR sampling yielded; Level 1 variable at the participant level; group-mean centered), (b) the total number of net hours of EAR monitoring (the number of waking and compliant hours over which the EAR sampling occurred; Level 1 variable at the participant level; group-mean centered), and (c) the proportion of EAR monitoring days that were weekend days (proportion of weekend days; expressed as a 0:1 ratio with 0 indicating weekday only [Monday to Friday] and 1 indicating weekend-only monitoring [Saturday/Sunday]; Level 2 variable at the sample level based on each study's EAR monitoring schedule).

These sensitivity analyses modeled each of these three methodological factors as a predictor of the outcome (i.e., WPD) and as a moderator of the effect of interest (i.e., the gender effect). We conclude that the methodological variable had an impact on the estimated gender difference if the 95% credible interval for the interaction effect excluded zero. In this case, we interpret the direction and magnitude of the effect through the effect size estimate (Cohen's d). If a methodological variable has a substantial zero-order effect but a minimal moderating effect, this implies that methodological factors affected the outcome (i.e., WPD) but did not bias the results of the key research questions (i.e., the effects of gender).

Deviations From the Preregistration

We implemented four analytic changes from the preregistration. All deviations from the preregistration/accepted Stage 1 article are described and justified in Table 3, which is based on the template by Willroth and Atherton (2024).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Deviations From the Preregistration/Accepted Stage 1 Protocol

No.		Detail	Deviation: Original wording	Deviation: Deviation description	Deviation: Reader impact
1	Type Reason Timing	Analysis Plan not possible After data access	RQ2: "Finally, we will test whether gender differences in words per day change across the age groups using a moderated regression model (with contrasts for the age group comparisons). Bayesian models allow for the coding and simulation of a parameter that represents a difference in differences, such as the difference in gender differences between adolescents and adults. These difference parameters will be created for all pairwise combinations of age cohorts and tested using the ROPE method." (p. 28, accepted Stage 1 article)	Following the preregistration, we ran the models separately for the four age groups: "For RQ2, we therefore estimate the gender difference separately for the four age groups. We then follow the procedure outlined for RQ1 to determine (a) whether a meaningful gender difference exists in each group (using the ±1,000-word ROPE), and (b) if so, what the magnitude of the estimated effect is (using Cohen's guidelines)" (p. 28, accepted Stage 1 article). We were unable to run the full model with contrasts for the age group comparisons. We did not manage to get the models to converge.	The deviation deprives the reader of knowledge of the extent to which the estimated gender differences differed credibly between the age groups. While such knowledge would be ideal, it appears not critical given the actual findings. Small gender differences comparable with the one reported by Mehl et al. (2007) emerged for three of the four age groups. A substantially larger gender difference emerged for middle adulthood. This difference was noticeably ("visibly") different from the other three (Figure 3). The age group comparisons were exploratory, and no hypothesis was preregistered.
2	Type Reason Timing	Analysis Typo/error After data access	RQ3: To what extent is experienced stress (as a marker of biobehavioral coping processes) associated with the gender difference in daily word use between men and women? Bayesian assessment of null values via ROPE analysis and Cohen's *d* estimates will be used to address this research question (Figure 2, accepted Stage 1 article).	We mistakenly proposed a ROPE approach (95% HDI using a ±1,000-word ROPE) to evaluate RQ3. The Stress × Gender interaction reflects how much a 1-point increase in POMP-scored stress changes the gender difference in WPD. We ultimately evaluated the extent to which stress had a moderating effect using the magnitude of the β weights (e.g., 11 WPD), along with the 95% credible interval and, as preregistered, the effect size (Cohen's *d*).	The deviation should not affect readers' interpretation of the results because the analyses and reported statistical information are identical to what was preregistered. We mistakenly "copied over" the decision criterion "±1,000 words" from the prior aims, not realizing that, testing an interaction with (rather than main effect of) gender, the β weight reflects a different metric. A "±1,000-word" effect of stress would be unduly large.
3	Type Reason Timing	Analysis Typo/error After data access	RQ4: How do the gender difference effects compare for objectively observed versus subjectively rated general talkativeness? Bayesian assessment of null values via ROPE analysis and Cohen's *d* estimates will be used to address this research question (Figure 2, accepted Stage 1 article).	We mistakenly proposed a ROPE approach (95% HDI using a ±1,000-word ROPE) to evaluate RQ4. The dependent variable is self-rated talkativeness, measured on a POMP metric. We ultimately evaluated the magnitude of the gender difference in subjectively rated general talkativeness using the magnitude of the β weights (e.g., 5.95), along with the 95% credible interval and, as preregistered, the effect size (Cohen's *d*).	The deviation should not affect readers' interpretation of the results because the analyses and reported statistical information are identical to what was preregistered. As above, we mistakenly "copied over" the decision criterion "±1,000 words" from the prior aims, not realizing that the dependent variable here has a POMP, not a WPD metric.
4	Type Reason Timing	Analysis Plan not possible After data access	"These sensitivity analyses will involve using each of the methodological factors listed as a covariate, and as a moderator of the effect of interest (e.g., the gender effect), in a series of separate models." (Figure 2, accepted Stage 1 article) The methodological factors listed were sampling frequency, length of one recording, the number of days over which the data were collected, the proportion of sampled days that were weekend days, the total recording time, and the number of hours over which the EAR monitoring occurred (p. 30 of the accepted Stage 1 article).	We made the following changes: • To be consistent with all other analyses, we used the 95% credible interval along with the effect size (Cohen's *d*) to evaluate the impact of a methodological factor (instead of the Bayes factor). • Several variables had minimal variability and discontinuous distributions that precluded linear analyses (e.g., most studies recorded 30 or 50 s; one study recorded 5 min); moreover, participants' actual EAR monitoring schedule often deviated substantially from the study's planned protocol. We therefore deemed the variables (a) total recording time, (b) number of hours over which	The deviations might affect readers' interpretation to the extent that they had concrete hypotheses about the impact of a specific factor (e.g., recording length). The deviations might strengthen the confidence of readers who thought that it is less ideal to analyze the different elements of the EAR sampling scheme (e.g., recording duration and frequency) as isolated variables and at the sample level and better to analyze them as composite variables and at the participant level (e.g., as amount of data available for each participant). We recommend that the results of the sensitivity analyses be interpreted with caution anyway because it is unfortunately clear that the data we had available for this project, although all the

(table continues)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3 (*continued*)

Deviation				
No.	Detail	Original wording	Deviation description	Reader impact
			the monitoring occurred, and (c) proportion of monitoring days that were weekend days best representing the methodological factor space. All three variables were preregistered and have sound distributional properties; the first two are computed using the information from all originally proposed factors.	data we found currently available in the scientific community, were insufficient for the Bayesian analyses to yield precise estimates.

Unregistered step					
No.	Detail		Original wording	Unregistered step description	Reader impact
1	Type Timing	Analysis After data access	"For gender differences, the difference coefficient will be tested." (Figure 2, accepted Stage 1 article) "Gender difference estimates for which the 95% HDI falls completely within a ±1,000 words ROPE centered around a zero difference will be interpreted as practically equivalent." (p. 27 of the accepted Stage 1 article)	Our preregistration failed to specify the centering of the categorical gender predictor. Given that gender varies within the sample (i.e., male and female participants) and between samples (i.e., the proportion of male vs. female participants in each sample), it must be modeled with two predictors that independently capture the within- and between-group effects. We ultimately used the UN(M) model (Yaremych et al., 2021) to statistically separate the within-sample (Level 1) and between-sample (Level 2) effects of gender.	This unregistered, corrective step should increase the readers' confidence in the results. The failure to center categorical predictors is common and one that must get more attention (Yaremych et al., 2021). Throughout the Stage 1 review process, we were unfortunately unaware of it. We thank Jessie Sun for bringing this issue to our attention and for sharing the article with us. The question whether women speak more WPD than men pertains to the within-sample effect; we therefore report the between-sample effects but do not interpret them.

Note. Adapted from "Best Laid Plans: A Guide to Reporting Preregistration Deviations," by E. C. Willroth and O. E. Atherton, 2024, *Advances in Methods and Practices in Psychological Science*, *7*(1). (https://doi.org/10.1177/25152459231213802). CC BY-NC-4.0. RQ = research question; ROPE = region of practical equivalence; POMP = percent of maximum possible; HDI = high-density interval; WPD = words per day; EAR = electronically activated recorder.

Results

Preliminary Descriptive Analyses: How Many Words Do Individuals Speak Every Day?

Based on the descriptives of the raw data (see Table 4), the 2,197 participants spoke on average an estimated 12,792 WPD (SD = 9,154), with an impressive range around this mean: The least talkative participant, an adult man, spoke 62 WPD, whereas the most talkative participant, also an adult man, spoke 124,134 WPD (range = 124,072 WPD). One additional female participant spoke more than 120,000 WPD (120,731), and two female and one male participant spoke more than 60,000 WPD (60,254; 67,000; 76,964). In sum, an effective range of <100 to >120,000 WPD is remarkable.

This compares with 15,959 WPD (SD = 7,949) with a minimum of 695 (male) and a maximum of 47,016 WPD (also male) among the 396 participants in the original Mehl et al. (2007) study (range = 46,321 WPD). The replication here therefore estimates the number of words individuals speak per day as lower than the original study (>3,000 WPD). Further, consistent with the larger sample size and more diverse sample composition, the replication finds a larger standard deviation (+ >1,000 WPD) and considerably wider range (+ >70,000 WPD).

RQ1: Is There a Gender Difference in Words Spoken per Day Between Men and Women?

Descriptives

The descriptive statistics for male and female participants in the full sample are provided in Table 4 and visualized in Figure 2. Men spoke on average 11,950 WPD (SD = 9,025), while women spoke on average 13,349 WPD (SD = 9,199). This compares with 15,660 WPD (SD = 8,633) for men and 16,215 (SD = 7,301) WPD for women in Mehl et al.'s (2007) study.

Statistical Test of RQ1

We used Bayesian multilevel models via the *brms* package in *R* (with four chains of 3,000 iterations and a warm-up of 1,000) to predict WPD from gender, with participants nested within each of our 22 samples. We modeled gender via two fixed effects, one at the within-sample level for the individual effect of gender and one at the between-sample level for the effect of sample gender composition, to separate within- and between-group effects of gender (UN(M) model; Yaremych et al., 2021). Theoretically, the question whether women speak more WPD than men is addressed by the within-sample effect. The between-sample effect indicates how much the gender composition of a sample influenced the WPD estimates. In other words, the between-sample effect shows the extent to which variability in the estimated gender difference is due to the proportions of females (or male) participants in samples deviating from parity (i.e., 50%), independent of the effect of gender at the individual (i.e., within-sample) level.

Table 4

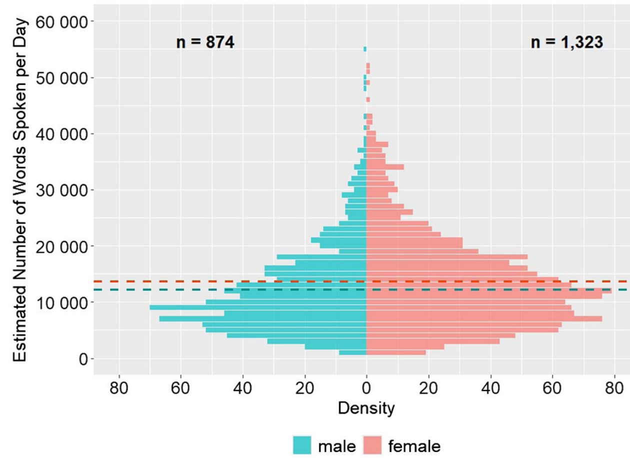
Descriptive Statistics for Research Question 1

Gender	Words spoken per day					Sample size
	M	Mdn	SD	Min	Max	N/n
All participants	12,792	11,013	9,154	62	124,134	2,197
Men	11,950	9,851	9,025	62	124,134	874
Women	13,349	11,620	9,199	143	120,731	1,323

Note. Min = Minimum; Max = Maximum.

Figure 2

Distribution of Estimated Number of Words Spoken per Day



Note. The distributions of the estimated number of words spoken per day (WPD) for the 874 male and 1,323 female participants in the analyses. The dashed lines indicate the mean values for men and women. Note that the descriptive (rather than model-implied) means are depicted here. The tests of the research questions report the model-implied means. The values of four participants with WPD > 60,000 are omitted for optimal display purposes. See the online article for the color version of this figure.

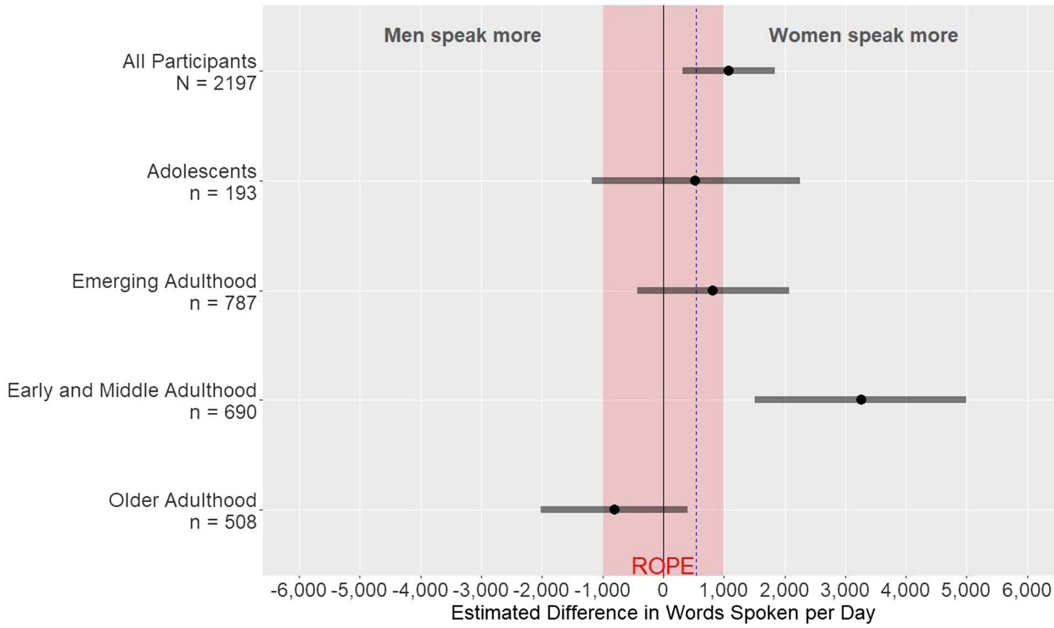
The estimated within-sample effect of gender was 1,073 WPD (95% CrI [316, 1,824]) indicating that female participants spoke on average 1,073 WPD more than male participants. The 95% credible interval includes substantial areas within and outside of our ROPE of 1,000 WPD, with the highest probability point estimate (1,073) falling just outside of it (see the first row of Figure 3). Our preregistered analysis plan specified that a conclusion of no practical difference required the full credible interval to fall within the 1,000 WPD ROPE. It further specified that a conclusion of the presence of a practical difference required the full credible interval to fall outside the 1,000 WPD ROPE. The results therefore provide ultimately—and unfortunately, despite the sample size of >2,000 participants, more than five-fold the original sample size—inconclusive evidence as there is neither sufficient statistical information to confidently conclude that women speak practically more WPD than men nor that the two genders speak a practically equivalent number of WPD. We do, though, have sufficient statistical information to conclude that men do not speak more WPD than women, as negative values are not credible parameter estimates. The estimated 1,073 WPD difference is about twice as large as the 546 WPD gender difference reported in the original study (Mehl et al., 2007).

Finally, we estimated the magnitude of the within-sample gender effect as Cohen's d = 0.13 (95% CI [0.04, 0.22]). Based on our preregistered analysis plan, this is interpreted as a small effect size.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 3

Estimated Gender Difference in Words Spoken per Day for All Participants and by Age Group



Note. (Within-sample) effects of gender on words spoken per day (WPD) for all participants (Research Question 1) and by age group (Research Question 2). The gray bars represent 95% credible intervals. The red-shaded area highlights the ±1,000 WPD ROPE. The dashed blue line marks the 546 WPD gender difference reported by Mehl et al. (2007). ROPE = regions of practical equivalence. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Looking at the means can provide greater context about the practical magnitude of this effect. Male participants spoke on average an estimated (i.e., model-implied) 11,950 WPD, while female participants spoke on average an estimated 13,349 WPD. Thus, the within-gender variability is roughly nine times as big as the difference between the two genders.

RQ2: To What Extent Does Age (As a Marker of Developmental Processes) Moderate a Gender Difference in Words Spoken per Day Between Men and Women?

Descriptives

Descriptive statistics for the four age groups (adolescence: 10–17 years; emerging adulthood: 18–24 years; early and middle adulthood: 25–64 years; older adulthood: ≥65 years) are summarized in Table 5 and visualized in Figures 3 and 4.

Based on the actual descriptive means (i.e., not the model-implied estimates), it appears that there were small gender differences in WPD among adolescent (women spoke 563 WPD more), emerging adult (women spoke 753 WPD more), and older adult (men spoke 508 WPD more) participants and a large gender difference in WPD among participants in early and middle adulthood (women spoke 2,069 WPD more). Nineteen participants did not provide their age. The gender difference among this group was also small (women spoke 370 WPD more). Data are visualized in Figure 4.

Statistical Test of RQ2

We used the same Bayesian multilevel modeling approach as in RQ1, again modeling the effect of gender at both the within-sample and between-sample levels. However, we now split the full data into four age group subsets and ran the analysis for each subgroup separately.

Adolescence. Among adolescent participants, the estimated within-sample effect of gender was 513 WPD (95% CrI [−1,206, 2,286]). This indicates that, in this age group, female participants

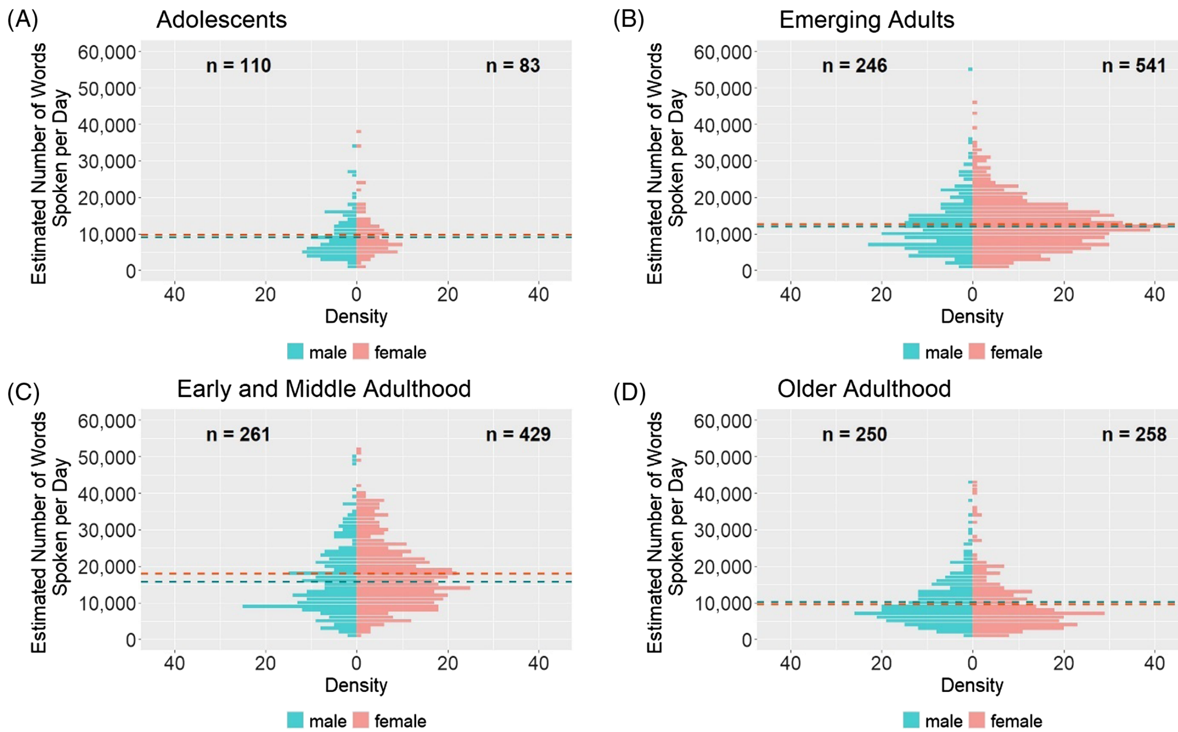
Table 5

Descriptive Statistics for Research Question 2

		Words spoken per day		Sample size
Age group	Gender	*M*	*SD*	*n*
Adolescence (10–17 years)	Men	8,635	5,903	110
	Women	9,198	6,298	83
Emerging adulthood (18–24 years)	Men	11,712	8,031	246
	Women	12,465	8,313	541
Early and middle adulthood (25–64 years)	Men	15,641	11,448	261
	Women	17,710	9,791	429
Older adulthood (≥65 years)	Men	9,709	6,577	250
	Women	9,201	7,597	258

Figure 4

Distribution of Estimated Number of Words Spoken per Day in the Four Age Groups: (A) Adolescents, (B) Emerging Adults, (C) Early and Middle Adulthood, and (D) Older Adulthood



Note. The distribution of the estimated number of words spoken per day for the 874 male and 1,323 female participants in the four age groups. Subsample sizes displayed on corresponding figures. The dashed lines indicate the mean values for men and women. Note that the actual descriptive (rather than model-implied) means are depicted here. The statistical tests of the research questions report the model-implied means. Participants with words spoken per day values >60,000 are omitted for optimal display purposes. See the online article for the color version of this figure.

spoke on average about 500 WPD more than male participants. The wide 95% credible interval (given the smaller subsample) includes values within and outside of the 1,000 WPD ROPE (second row of Figure 3). Therefore, while the point estimate suggests no practical gender difference, we do not have sufficient statistical information to conclude practical equivalence. The Cohen's *d* of the effect was 0.08 (95% CI [−0.20, 0.38]) suggesting a small effect size, very similar to the one estimated by the original study ($d = 0.07$).

The between-sample effect of gender was 31,894 WPD (95% CrI [−261,434, 350,519]) indicating that adolescent samples with a larger proportion of female participants had higher WPD estimates. The between-sample gender effect is not relevant for RQ2.

Emerging Adulthood. Among emerging adult participants, the estimated within-sample effect of gender was 841 WPD (95% CrI [−369, 2,028]). This indicates that, in this age group, women spoke on average a little over 800 WPD more than men. The 95% credible interval includes values within and outside of the 1,000 WPD ROPE (third row of Figure 3). Therefore, while the point estimate suggests no practical gender difference, we do not have sufficient statistical information to conclude practical equivalence. The Cohen's *d* of the effect was 0.11 (95% CI [−0.05, 0.26]) suggesting a small effect size, comparable with the one estimated by the original study ($d = 0.07$).

The between-sample effect of gender was −3,021 words (95% CrI [−17,198, 12,793]), indicating that emergent adult samples with a larger proportion of male participants had higher WPD estimates. The between-sample gender effect is not relevant for RQ2.

Early and Middle Adulthood. Among participants in early and middle adulthood, the estimated within-sample effect of gender was 3,275 WPD (95% CrI [1,492, 5,074]). This indicates that, in this age group, women spoke on average more than 3,000 WPD more than men. The 95% credible interval falls fully outside the 1,000 WPD ROPE (fourth row of Figure 3). Therefore, in this age group, women speak practically more WPD than men. Cohen's *d* of the effect was 0.32 (95% CI [0.14, 0.49]) suggesting a small to medium effect size, roughly four times the one estimated by the original study ($d = 0.07$). Looking at the estimated means, in this age group, men spoke on average 18,570 WPD, while women spoke on average 21,845 WPD.

The between-sample effect of gender was −6,628 words (95% CrI [−12,725, −462]), indicating that early and middle adulthood samples with a larger proportion of male participants had higher WPD estimates. The between-sample gender effect is not relevant for RQ2.

Older Adulthood. Among older adult participants, the estimated within-sample effect of gender was −788 WPD (95% CrI [−2,013, 417]). This indicates that, in this age group, women spoke on average about 800 WPD *less* than men. The 95% credible interval includes values both within and outside of the 1,000 WPD ROPE (third row of Figure 3). Therefore, while the point estimate suggests no practical gender difference, we do not have sufficient statistical information to

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

conclude practical equivalence. Cohen's d of the effect was −0.11 (95% CI [−0.29, 0.06]), suggesting a small effect size, in this case in the direction of men speaking more WPD than women.

The between-sample effect of gender was 4,090 words (95% CrI [16,810, 25,580]), indicating that older adult samples with a larger proportion of female participants had higher WPD estimates. The between-sample gender effect is not relevant for RQ2.

RQ3: To What Extent Does Experienced Stress (As a Marker of Biobehavioral Coping Processes) Moderate a Gender Difference in Words Spoken per Day Between Men and Women?

We evaluated RQ3 with a Bayesian multilevel model that had gender as within- and between-sample predictor (UN(M) model), stress as within- and between-sample predictor (UN(M) model), and the interaction term between within-sample gender and within-sample stress, in the subsample of participants who had a measure of experienced stress (n = 1,227). The stress measure was POMP scored.

For the test of RQ3, the interaction effect was the only effect of interest. The within-sample Gender × Stress interaction was estimated as 11 WPD (95% CrI [−46, 68]) and a Cohen's d of 0.001 (95% CI [−0.006, 0.009]). Based on the minimal effect size, the close-to-zero estimated WPD difference, and the credible interval including negative and positive values, we conclude that experienced stress had no measurable effect on the gender difference in WPD.

Beyond relevance for RQ3 (and beyond the preregistration), it was interesting that the estimated within-sample effect of stress was −44 WPD (95% CrI [−93, 4]), indicating that for every 1-point increase in (POMP-scored) stress, participants spoke on average 44 fewer WPD. The magnitude of this effect was very small, Cohen's d = −0.006; 95% CI [−0.01, 0.0005], although it amounts to approximately a 1,500 WPD difference between a person 1 SD below and 1 SD above the mean.

RQ4: How Do Gender Differences Compare for Objectively Observed Versus Subjectively Rated General Talkativeness?

We addressed RQ4 via a Bayesian multilevel model like the one in RQ1, except with self-rated general talkativeness replacing objectively observed talkativeness (i.e., WPD).

Overall Gender Difference (RQ1)

For the full sample of participants with a self-rated talkativeness score (n = 1,227), the model estimated that male participants rated their talkativeness (POMP-scored) as 52.08 (intercept), with female participants rating themselves as 5.95 POMP points more talkative (within-sample gender effect; 95% CrI [2.84, 8.92]). The magnitude of this effect, d = 0.23 (95% CI [0.11, 0.34]), is small to medium and comparable with the corresponding effect for observed talkativeness (d = 0.15, 95% CI [0.06, 0.24]).

Gender Differences in the Age Groups (RQ2)

We further estimated gender differences in self-rated talkativeness for each age group. No adolescent participant had self-rated talkativeness data, so we could only estimate models for emerging, early and middle, and older adulthood.

For emerging adulthood (n = 422), the model estimated that male participants rated their talkativeness as 73.65, with female participants rating themselves as 9.94 POMP points more talkative (95% CrI [4.52, 15.36]). The magnitude of this effect, d = 0.38 (95% CI [0.17, 0.59]), is considerably larger than the corresponding effect for observed talkativeness (d = 0.11, 95% CI [−0.05, 0.26]).

For early and middle adulthood (n = 424), the model estimated that male participants rated their talkativeness as 49.82, with female participants rating themselves as 5.32 POMP points more talkative (95% CrI [−0.18, 10.75]). The magnitude of this effect, d = 0.23 (95% CI [0.11, 0.35]), is small to medium and comparable with the corresponding effect for observed talkativeness (d = 0.32, 95% CI [0.14, 0.49]).

For older adulthood (n = 369), the model estimated that male participants rated their talkativeness as 53.43, with female participants rating themselves as 3.19 POMP points more talkative (95% CrI [−2.25, 8.66]). The magnitude of this effect, d = 0.12 (95% CI [−0.08, 0.33]), is small and comparable with the corresponding effect for observed talkativeness but in the opposite direction (d = −0.11, 95% CI [−0.29, 0.06]).

Moderating Effect of Experienced Stress (RQ3)

Last, we estimated the extent to which experienced stress moderated the within-sample gender effect for self-rated talkativeness. The model estimated the interaction between gender and stress as 0.15 (95% CrI [−0.12, 0.42]). The magnitude of this effect, d = 0.006 (95% CI [0.005, 0.017]), is minimal and comparable with the corresponding interaction effect for observed talkativeness (d = 0.001, 95% CI [−0.006, 0.009]).

Exploratory Analyses Beyond Those Preregistered Within the Stage 1 Report

One unexpected aspect of the preliminary descriptive analyses that caught our interest was that the present study estimated the number of words spoken per day at about 3,000 words lower than the original study (M_{present} = 12,792 vs. M_{original} = 15,959). As an additional analysis beyond the preregistration, we therefore explored the extent to which WPD may have decreased over time, that is, as a linear function of the year in which the study was run. For this, we reran the Bayesian multilevel model for RQ1 with study year (measured as the difference between the year in which data collection for a sample was started and 2005, the year of data collection for the oldest included sample) as a main effect. In 2005, participants spoke an estimated 16,632 WPD (95% CrI [13,545, 19,780]). The effect of study year was −338 WPD (95% CrI [−652, −25]), indicating that, for every additional year between 2005 and 2018, participants spoke about 300 fewer WPD. The magnitude of this effect per year was very small, d = −0.04 (95% CI [−0.08, −0.003]). However, a decrease of more than 3,000 WPD over a decade, if robust, would be nontrivial.

Sensitivity Analyses

To explore the extent to which differences in EAR sampling procedures between the 22 samples accounted for the estimated gender difference in WPD, we tested three methodological variables related to the quantity and context of the monitoring: (a) the total recording time (the net awake and compliant number of minutes of recording that the EAR sampling yielded, Level 1 variable at the participant level, group-mean centered), (b) the total number of net hours of EAR monitoring (the number of waking and compliant hours over which the EAR sampling occurred, Level 1 variable at the participant level, group-mean centered), and (c) the proportion of EAR monitoring days that were weekend days (proportion of weekend days, expressed as a 0–1 ratio with 0 indicating weekday-only [Monday to Friday] and 1 indicating weekend-only monitoring [Saturday/Sunday], Level 2 variable at the sample level based on each study's EAR monitoring schedule).

For RQ1 and RQ2, we ran two models for each of the three variables, a predictor-only model to test for the zero-order effect of the methodological variable on the dependent variable, WPD, along with the zero-order effect of within-sample gender, and an interaction model, which included the interaction term between within-sample gender and the methodological variable. For RQ3, we ran only one model that included the predictors within-sample gender, POMP scored stress, and the methodological variable with all main effects and interactions (because the target effect was an interaction). For all three research questions, we conclude that the methodological variable had an impact on the estimated gender difference in WPD if the 95% HDI for the target interaction effect excluded zero. In such cases, we then interpret the direction and magnitude of the effect through the effect size estimate (Cohen's d). No sensitivity analyses were conducted for RQ4 because the analyses there reestimated all RQ1–3 effects with self-reported talkativeness as dependent variable, which was not the focus of our analyses.

The results of the full sensitivity analyses, along with the data and code to reproduce them, are available on the Open Science Framework. For space reasons, we report here a concise summary along with all analyses that yielded credible evidence for a methodological effect on the research questions.

RQ1

The sensitivity analyses for the full sample provided no evidence that any of the three methodological variables had a credible effect on the magnitude of the estimated gender difference in WPD. The 95% HDIs for all interaction effects contained zero as a plausible value.

RQ2

The sensitivity analyses for three of the four age groups, adolescence, emerging adulthood, and middle adulthood, provided no evidence that any of the three methodological variables had a credible effect on the magnitude of the estimated gender difference in WPD. The 95% HDIs for all interaction effects contained zero as a plausible value. Among older adults, the sensitivity analyses suggested that those who had more available EAR data and more hours of EAR monitoring had a (minimally) smaller estimated gender difference in WPD (recall that the gender difference in this age group was that men spoke more WPD than women). The 95% HDIs for these two interaction effects excluded zero as a plausible value, and the estimated effect sizes were very small ($d = -0.004$ and $d = -0.02$, respectively). For the third methodological variable, the sensitivity analyses suggested that older adult participants who had a higher proportion of EAR monitoring over the weekend had a (much) smaller estimated gender difference in WPD. The 95% HDI for this interaction effect excluded zero as a plausible value, and the estimated effect size was very large ($d = -3.37$). We have no good explanation for this potential methodological effect but highlight that the preregistered analyses did not yield a large gender difference for this age group to begin with (−788 WPD, 95% CrI [−2,013, 417]; $d = -0.11$, 95% CI [−0.29, 0.06]).

RQ3

The sensitivity analyses provided no evidence that any of the three methodological variables had a credible effect on the magnitude of the effect of stress on the estimated gender difference in WPD. The 95% HDIs for all Methodological Variable × Within-Sample Gender × POMP-Scored Stress interaction effects contained zero as a plausible value.

Taken together, the sensitivity analyses that we were able to conduct provide little evidence of systematic methodological influences related to the EAR sampling on the findings. However, for several analyses, particularly the analyses of age subgroups, the limited amount of available data (i.e., small subsamples) resulted in high uncertainty of the models and estimates. Moreover, the methodological variable *proportion of weekend monitoring* had limited variability (i.e., the studies ultimately did not differ very much in their EAR sampling protocols), particularly for the age subgroup analyses, which also resulted in high uncertainty of the models and estimates. Therefore, we consider these sensitivity analyses adding some support for the validity of our results rather than "clearing" them from methodological artifacts or biases. Just like with the main analyses, although this study used all EAR data that we found currently available in the scientific community, it is unfortunately ultimately not enough for precise Bayesian estimates.

Discussion

The main aim of this registered replication study was to replicate Mehl et al.'s (2007) study (Are Women Really More Talkative than Men?) by reestimating the number of words that men and women speak in a day and reevaluating the magnitude of the gender difference using a new (i.e., nonoverlapping with the original study), large data set of 2,197 participants (more than five times the original sample size) and 631,030 ambient sound recordings (pooled over 22 samples). Beyond this main aim, we sought to explore the extent to which age, as a marker of developmental processes, and experienced stress, as a marker of biobehavioral coping processes, are associated with this gender difference. Finally, we sought to compare the general, age-, and stress-related gender differences for objectively observed talkativeness with those for subjectively rated talkativeness.

At the broadest level, the study confronted us with the (disappointing) finding that, despite the large sample size and our effort to gather and use all existing data (at the time) for addressing these questions, all but one of the analyses yielded ultimately inconclusive evidence. The data provided insufficient statistical information to conclude practical equivalence, that is, that the two genders speak a practically equivalent number of WPD, or practical nonequivalence, that is, that either gender speaks practically more WPD than the other. Because we sought to replicate the absence of a widely assumed gender difference, we employed Bayesian analyses to allow for a direct test of the null.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Moreover, because self-replications need tight decision criteria, we chose <1,000 WPD as a threshold for an effectively meaningless gender difference. Our decision rule thus was whether *the full 95% credible interval* would fall within versus outside of the ± 1,000 WPD ROPE (Kruschke, 2018).

In our only confirmatory test, the test in the full sample (N = 2,197) for which we hypothesized no gender difference, the width of the credible interval was 1,508 WPD (i.e., ± 754 WPD). This means that our statistical precision effectively limited us to considering (maximum probability) gender difference estimates of <246 WPD practically equivalent (246 + 754 = 1,000 WPD), which is less than half of the original study's estimate (546 WPD). Ironically, this leads to the awkward scenario where evidence identical to (or even substantially smaller than) the original estimate would have been deemed inconclusive here. Said differently, even though this study had more than five times the number of participants compared with the original one, its analyses convey a lot more uncertainty than the original study portrayed. This acknowledgment of large statistical uncertainty, as humbling as it is for this registered replication, is consistent with the field's emerging understanding of what (often surprisingly large) sample sizes are needed to achieve robust and generalizable effects (Yarkoni, 2022).

Importantly, the widths of the credible intervals for the other inconclusive tests were even larger, given that they are derived from subsamples (range = 2,397 WPD for emerging adults, n = 787; 3,492 WPD for adolescents, n = 193). In addition, the only test that did yield conclusive evidence—the test for a gender difference in early and middle adulthood (ages 24–65; n = 690)—yielded a 3,582 WPD wide credible interval with women speaking more WPD than men, where the upper bound (5,074 WPD) would suggest a very large and the lower bound (1,492 WPD) only a modest gender difference. Therefore, at the most zoomed-out level, this study finds, in effect, that even with the best faith effort to gather and use all existing data to evaluate a research question, we often do not have the statistical precision we would need to come to unambiguous conclusions. Considering that this study relied on data that were gathered with the support of many grants, collected over a period of 13 years, and transcribed by hundreds of research assistants in tens of thousands of hours, this (painful) realization is important to "sit with." With the background of this acknowledged large statistical uncertainty, what can this registered replication contribute to scientific knowledge of gender differences in everyday talkativeness?

Is There a Gender Difference in WPD Between Men and Women?

Regarding the overall gender difference (RQ1), where we expected to replicate Mehl et al.'s (2007) finding of no (practically important) difference, we can, with statistical confidence, rule out the possibility that men speak more WPD than women. This is important because a comprehensive meta-analysis by Leaper and Ayres (2007) found (counter to their initial prediction) men to be more talkative than women. Importantly, however, this meta-analysis identified effect size heterogeneity that, at a closer look, aligns pertinent subfindings better with the results of this study. Specifically, it estimated close-to-zero differences (d = 0.01 and d = −0.03) for talkativeness operationalized as the number of words spoken and for data collected outside the lab. In this context, it is important that our study, due to limitations around wearing the EAR at work, heavily oversampled conversations outside of the workplace, thereby underrepresenting specific (e.g., agentic and noncollaborative) social contexts in which men have been theoretically predicted and empirically shown to outtalk women (Leaper & Ayres, 2007; Onnela et al., 2014).

Our analyses further rule out that overall, averaging over all age groups, a zero difference in WPD is a plausible value. Said differently, at the most zoomed-out level, our study finds that women overall speak more words per day than men, at least when studied across the contexts that the EAR can representatively sample. The maximum probability estimate for this difference was 1,073 WPD, about twice as large as the 546 WPD gender difference reported in the original study and just slightly larger than our 1,000 WPD ROPE. Therefore, this overall finding (RQ1) updates the knowledge from Mehl et al.'s (2007) study that women are, to some extent, more talkative. Notably, though, the within-gender variability was roughly nine times as big as the estimated difference between the two genders. Regarding the magnitude, the credible interval shows that a gender difference as small as 316 WPD (clearly trivial) or as large as 1,824 WPD (potentially meaningful) is ultimately plausible given the data, thereby rendering the test of our preregistered prediction inconclusive.

How Does Age Matter for the Gender Difference in WPD Between Men and Women?

At the finer grained level, our study yielded interesting exploratory findings about how age, as a marker of developmental processes, might matter for the gender difference in WPD (RQ2). Because the age group analyses relied on much smaller samples, only for early and middle adulthood (a single age group, ages 24–65) did we have enough statistical information to draw a conclusion based on our ROPE criterion. For the three other age groups, we unfortunately could not confidently distinguish between practical equivalence and a practically important gender difference.

Based on the maximum probability parameter estimates, women tend to speak about 500 and 800 WPD more than men in adolescence (10–17 years) and emerging adulthood (18–24 years), respectively. These numbers, and corresponding effect sizes (d = 0.07 and d = 0.11), are broadly consistent with—and, in fact, quite close to—the ones reported by Mehl et al. (2007), which are based on a college student sample (546 WPD; d = 0.07). From a broader replicability perspective, then, it is notable that this registered replication, while not confirming the preregistered prediction across the full sample, does replicate the original gender difference quite closely in its estimates for participants of comparable ages. Again, however, the wide credible intervals indicate that both rather small and quite large population values are plausible, thereby rendering the equivalence test based on our ROPE criterion inconclusive.

Interestingly, for participants in early and middle adulthood (25–64 years), this study yielded a maximum probability parameter estimate of more than 3,000 WPD (d = 0.32). This effect is more than six times larger than the gender difference reported by Mehl et al. (2007). It is consistent with the societal stereotype that women talk more than men, as well as the recent finding that women tend to write more words than men in a narrative writing task (d = 0.31; Schultheiss et al., 2021). The credible interval for the 25–64-year age group was again wide (95% CrI [1,492, 5,074]); however, it excluded all values falling within the 1,000 WPD ROPE. We can conclude that men and women in this age group do not speak a practically equivalent number of WPD. This clear gender difference in early and middle adulthood,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

although not predicted, is an important exploratory finding and should be considered a critical update to the scientific knowledge of gender differences in everyday talkativeness.

Finally, among older adults, the maximum probability parameter estimate suggests that men speak about 800 words more per day than women. We caution against an interpretation of this apparent "sign flip," given that the credible interval includes zero. Interestingly, this estimate appears to render generational explanations for the results in the other age groups, such as a fading of traditional gender-role socialization and corresponding gain of gender equality over historical time, unlikely. Such explanations would seem to require negatively graded effect size trends from older to younger or earlier to later-born participant groups, a pattern that is inconsistent with the estimate for older adults. Also undermining such a generational explanation, the emerging adult participants in Mehl et al.'s (2007) studies would now, 10+ years later, all fall into the early and middle adulthood category. Given that they did not show a substantial gender difference back then, a large gender difference suddenly emerging for them in early adulthood goes beyond a simple generational socialization perspective and would at least require an interactionist perspective. Last, the inconsistency of our data with a gain-of-gender-equality-over-historical-time explanation aligns with the recent finding that, while gender stereotypes have changed over the past 70 years, they have not consistently moved toward gender equality (Eagly et al., 2020).

An important question that emerges from our study, then, concerns what factor(s) might explain why women tend to speak more words than men particularly in early and middle adulthood. Potential explanations might revolve around underlying biological factors, such as sex hormones (e.g., estradiol) linked to verbal fluency advantages for women relative to men (Schultheiss et al., 2021), which should predominantly manifest or be accentuated between puberty and menopause (although the absence of a pronounced gender difference among emerging adult participants appears inconsistent with such an explanation). Other potential explanations might revolve around underlying sociocultural factors, such as traditional gender-role expectations that tend to afford women a greater responsibility in the communal domains of child rearing and family care (Eagly et al., 2020), which should also predominantly manifest (or be accentuated) in this age range. That is, it seems plausible that the gender difference could be partly explained by women talking to their children and other care dependents more than men do.

In this context, it is again important to highlight that there are inherent (ethical and legal) limitations around wearing the EAR at work. This study's database thus critically underrepresents workplace conversations and overrepresents leisurely and family conversations, rendering the obtained findings likely less representative of agentic and more representative of communal conversation contexts. Importantly, though, both the workplace and the leisure and family environment afford agentic and communal (conversation) behavior, just to different degrees (e.g., Onnela et al., 2014). Consistent with the idea that women might particularly speak more words than men in early and middle adulthood because of their stronger engagement in child rearing and family care, prior EAR studies on parent–child interactions have documented relatively strong gender-linked, and gender-role-consistent, communication patterns, particularly in the context of parental care (e.g., Alisic et al., 2017; Mangelsdorf et al., 2019).

Of course, other biological, sociocultural, and interactionist explanations are conceivable (see Eagly & Revelle, 2022, for a recent discussion). Ultimately, it is important to recognize that this study was not designed to test, and is therefore not in the position to speak to, the validity of different causal explanations. Systematic experimental approaches (which test specific theoretical hypotheses, e.g., Galinsky et al., 2024) and large-scale research syntheses (e.g., Leaper & Ayres, 2007) are in a better position to accomplish this. On the background of the original study (Mehl et al., 2007) being in response to postulated (large) brain-based sex differences in talkativeness (Brizendine, 2007), however, we do feel that the patterning of findings in this replication permits ruling out such an explanation for the number of words women and men speak every day. Such an explanation would appear to require either a uniform, substantial WPD gender difference across the full studied age range (if the innate brain-based sex differences are assumed to manifest early in development) or a substantial WPD gender difference emerging in adulthood and continuing into old age (if the innate brain-based sex differences are assumed to manifest only upon full brain maturation). The distinct lack of evidence regarding women speaking more WPD than men among the (cognitively healthy) older adult participants ($n = 507$) is clearly inconsistent with such an explanation.

How Does Stress Matter for the Gender Difference in WPD Between Men and Women?

We further evaluated to what extent experienced stress, as a marker of biobehavioral coping process, matters for the WPD gender difference (RQ3). Following the logic of Taylor et al.'s (2000) tend-and-befriend model, according to which women are more likely than men to respond to stress with affiliation, the WPD gender difference might be larger at higher levels of distress. Among the 966 participants for whom a stress measure was available, we found little evidence for that. A 1 percentage point increase in stress was associated with only an 11 WPD increase ($d = 0.001$). As there are many ways for an increased affiliative tendency to manifest in social behavior, our null finding has limited bearing on the validity of the tend-and-befriend model. However, we can conclude with reasonable confidence that gender differences in everyday talkativeness are unlikely to be exacerbated by stress.

Incidentally, and beyond the aims of this study, we found that stress negatively predicted WPD (for both genders). Specifically, a 1 percentage point increase in stress was related to a decrease of 44 WPD. Although the effect size was very small ($d = -.006$) and plausibly null (the credible interval spanned positive and negative values), this amounts to approximately a 1,500 WPD difference between a person 1 *SD* below and 1 *SD* above the mean, one and a half times as much as the estimated overall gender difference (1,073 WPD). If robust, such an effect would be consistent with the idea that stress can undermine social connection, thereby ironically undercutting the availability of social support when it is most needed.

How Do Gender Differences Compare for Objectively Observed Versus Self-Rated Talkativeness?

For a subsample of 1,227 participants, subjectively rated talkativeness was available from the Big Five Inventory item "I consider myself to be a person who is talkative." This allowed comparing the obtained gender differences in WPD with self-report estimates. The idea that

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

guided this comparison was that talkativeness can look different from the inside than from the outside (Vazire, 2010) and that the wide societal availability of the stereotype of female talkativeness might accentuate the gender difference from the perspective of the self. Across the full sample, the WPD and self-reported talkativeness measures were modestly correlated $r = .22$ (95% CI [0.17, 0.27]). Similar patterns of findings emerged for the overall gender difference, the gender difference in early and middle adulthood, and the effect of stress on the gender difference when using either self-reported or objective measures. Among emerging adult participants, however, a considerably (more than three times) larger gender difference emerged in self-reported talkativeness relative to WPD, and among older adult participants, women rated themselves somewhat more talkative than men even though no such gender difference emerged when using the WPD measure (if anything, older adult men descriptively spoke slightly more WPD than older adult women). Overall, then, no clear (e.g., accentuated) pattern emerged with respect to inside versus outside perspectives on talkativeness, although, from the perspective of the self, women generally perceived themselves in line with the stereotype (i.e., as more talkative than men), whereas, from a daily spoken word count perspective, that was not the case for older adults.

Limitations, Constraints on Generality, and Future Directions

The findings from this study are subject to important limitations that ultimately affect their reliability and constrain their generalizability. Most directly and perhaps most importantly, even though this study collected and analyzed all available EAR data ($N = 2,197$; more than five times the sample size of the original study), the Bayesian ROPE analyses revealed that the findings carry large statistical uncertainty (i.e., wide credible intervals). This statistical uncertainty, combined with our tight preregistered ±1,000 WPD ROPE criterion, prevented a conclusive test of whether men and women speak a practically equivalent number of words per day. To the extent that the research question is deemed important enough, future research could update the findings obtained here when/if sufficient data are available, such as from ongoing EAR studies and/or other suitable methods, to permit more precise effect estimates. Alternatively, future research could look at the existing data through different statistical lenses, such as opting for 66%, rather than our preregistered 95%, credible intervals (Kruschke, 2018) or arguing that only larger differences, say, exceeding 2,000 WPD, practically matter. In this spirit, we provide, on the Open Science Framework, an expanded summary figure that simultaneously shows 95% and 66% credible intervals and 1,000 WPD and 2,000 WPD ROPEs.

The generalizability of the obtained findings is further limited by an important lack of diversity/representation in the pooled sample, notably with respect (but not limited) to country of origin (data from only four different countries were included), sociocultural background (including racial/ethnic identification and socioeconomic status), and sexual orientation and gender identity (Patterson et al., 2004; Tornello, 2020). Gender roles, and associated behavioral norms, can vary widely across these elements, and it is therefore conceivable, if not likely, that gender differences in daily word use vary as a function of (some of) them. Moreover, because our focus was on the general talkativeness stereotype, this study did not investigate how aspects of the social context (e.g., gender composition of a group; agentic vs. communal affordances of the setting) can systematically affect how many words men or women speak in a certain (type of) context. In that regard, it is important to reiterate that the EAR studies analyzed here did not—and, at least in part, could not—sample workplace conversations, thereby rendering the findings less representative of agentic and more representative of communal conversation contexts. To the extent that women talk more than men particularly in communal contexts, smaller (or even reversed) gender differences might result when agentic contexts are representatively captured (Leaper & Ayres, 2007). As discussed above, it is possible that the unexpectedly large gender difference in early and middle adulthood may, in part, be the result of men and women, being differentially assorted into social contexts that maximally differ in communion during this developmental period.

Last, this study focused exclusively on gender differences in daily spoken words (in-person or over the phone). We did not consider how the gender difference might vary as a function of the social contexts the participants were in. Within the context elements that studies tend to code from the EAR sound files, the gender of the conversation partner (e.g., Badura et al., 2018; Karpowitz & Mendelberg, 2014), as well as the conversational setting, such as talking to a child or a romantic partner, being in a professional/work environment (which is a context the EAR selectively undersampled due to privacy regulations), and being in a private or public setting, would be theoretically potentially interesting variables. While we acknowledge that these contexts are likely to affect gender differences in daily word count, exploring them here was beyond the scope of this (replication) project. Future research could address the important question of context variations in words spoken per day, especially in early and middle adulthood, where a divergence in roles between women and men related to child-rearing responsibilities might be most pronounced.

Furthermore, with most individuals, at least in the countries studied here, owning a smartphone and computer-mediated communication, including email, text messaging, and social media, have become increasingly popular and are by now highly prevalent and for some possibly even dominant, communication mediums. Naturally, gender differences in "digital talkativeness" can differ from the estimates obtained for the spoken word here. Mobile sensing methods, which allow for a comprehensive (close to) "360 assessment" of a person's daily spoken and digital interactions, provide the opportunity to assess this possibility (Harari et al., 2020; Roos et al., 2023).

On this topic, we want to highlight an intriguing incidental "side finding" that emerged in exploratory analyses. Whereas the original study estimated people's daily spoken word use at around 16,000 WPD, the present study, using the same methods, estimated that number at roughly 3,000 words lower, around 13,000 WPD. Furthermore, we found that participants spoke roughly 300 WPD less every year between 2005 (the year the earliest sample was collected) and 2018 (the year the most recent sample was collected), resulting in an estimated "loss" of more than 3,000 spoken WPD over a decade. This effect was not preregistered, so should be interpreted with caution. Furthermore, we have no means to disambiguate causal factors behind this (possible) reduction in daily spoken words in this study. However, the dramatic rise of digital forms of communication emerges as a clear candidate explanation. If this reduction in daily spoken words indeed represents a loss of spoken communication to digital communication, then this study

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

would be among the first to quantify this communication shift using an intuitive real-world metric.

Summary and Conclusion

Women are widely assumed to be more talkative than men. The purpose of this study was to conduct a registered replication and extension of Mehl et al.'s (2007) study, which first found only a trivial difference in men's and women's daily spoken word use among college students. The present study addresses concerns about the original study's generalizability beyond college students and to different age groups. Across 2,197 (new) participants—more than five-fold the original sample size—men spoke on average 11,950 WPD and women 13,349 WPD, with very large individual differences (the least talkative participant spoke fewer than 100 WPD, the most talkative more than 120,000 WPD). The estimated gender difference (1,073 WPD; d = 0.13) was about twice as large as in the original study (546 WPD; d = 0.07). Smaller differences emerged among adolescent (513 WPD; $d = 0.08$), emerging adult (841 WPD, $d = 0.11$), and older adult (−788 WPD; $d = -0.11$) participants, but a substantially larger difference emerged for participants in early and middle adulthood (3,275 WPD; $d = 0.32$). Unfortunately, though, despite the considerable sample size(s), all parameter estimates carried large statistical uncertainty and, except for the gender difference in early and middle adulthood, provide inconclusive evidence regarding whether (on the basis of the preregistered ±1,000 WPD ROPE criterion) the two (binary) genders ultimately differ in a practically meaningful way in how many words they speak on a daily basis.

References

Abel, D. B., Salyers, M. P., Wu, W., Monette, M. A., & Minor, K. S. (2021). Quality versus quantity: Determining real-world social functioning deficits in schizophrenia. *Psychiatry Research*, *301*, Article 113980. https://doi.org/10.1016/j.psychres.2021.113980

Alisic, E., Barrett, A., Bowles, P., Babl, F. E., Conroy, R., McClure, R. J., Anderson, V., & Mehl, M. R. (2015). Ear for recovery: Protocol for a prospective study on parent–child communication and psychological recovery after paediatric injury. *BMJ Open*, *5*(2), Article e007393. https://doi.org/10.1136/bmjopen-2014-007393

Alisic, E., Gunaratnam, S., Barrett, A., Conroy, R., Jowett, H., Bressan, S., Babl, F. E., McClure, R., Anderson, V., & Mehl, M. R. (2017). Injury talk: Spontaneous parent–child conversations in the aftermath of a potentially traumatic event. *Evidence-Based Mental Health*, *20*(4), e19–e20. https://doi.org/10.1136/eb-2017-102736

Badura, K. L., Grijalva, E., Newman, D. A., Yan, T. T., & Jeon, G. (2018). Gender and leadership emergence: A meta-analysis and explanatory model. *Personnel Psychology*, *71*(3), 335–367. https://doi.org/10.1111/peps.12266

Beer, A., & Vazire, S. (2017). Evaluating the predictive validity of personality trait judgments using a naturalistic behavioral criterion: A preliminary test of the self-other knowledge asymmetry model. *Journal of Research in Personality*, *70*, 107–121. https://doi.org/10.1016/j.jrp.2017.06.004

Bierstetel, S. J., & Slatcher, R. B. (2020). Couples' behavior during conflict in the lab and diurnal cortisol patterns in daily life. *Psychoneuroendocrinology*, *115*, Article 104633. https://doi.org/10.1016/j.psyneuen.2020.104633

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

Brizendine, L. (2007). *The female brain* (1st ed.). Morgan Road Brooks.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. https://doi.org/10.1037/h0047470

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press. https://doi.org/10.1525/9780520350519

Calabrese, W. R., Emery, L. T., Evans, C. M., & Simms, L. J. (2024). *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition, personality disorders and the alternative model: Prediction of naturalistically observed behavior, interpersonal functioning, and psychiatric symptoms, 1 year later. *Personality Disorders: Theory, Research, and Treatment*, *15*(5), 361–370. https://doi.org/10.1037/per0000677

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346. https://doi.org/10.1207/S15327906MBR3403_2

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*(4), 385–396. https://doi.org/10.2307/2136404

Czopp, A. M., Kay, A. C., & Cheryan, S. (2015). Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, *10*(4), 451–463. https://doi.org/10.1177/1745691615588091

Donovan, R. A. (2011). Tough or tender: (Dis)similarities in White college students' perceptions of Black and White women. *Psychology of Women Quarterly*, *35*(3), 458–468. https://doi.org/10.1177/0361684311406874

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*(3), 301–315. https://doi.org/10.1037/amp0000494

Eagly, A. H., & Revelle, W. (2022). Understanding the magnitude of psychological differences between women and men requires seeing the forest and the trees. *Perspectives on Psychological Science*, *17*(5), 1339–1358. https://doi.org/10.1177/17456916211046006

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Lawrence Erlbaum.

Farrell, A. K., Slatcher, R. B., Tobin, E. T., Imami, L., Wildman, D. E., Luca, F., & Zilioli, S. (2018). Socioeconomic status, family negative emotional climate, and anti-inflammatory gene expression among youth with asthma. *Psychoneuroendocrinology*, *91*, 62–67. https://doi.org/10.1016/j.psyneuen.2018.02.011

Fingerman, K. L., Huo, M., Charles, S. T., & Umberson, D. J. (2020). Variety is the spice of late life: Social integration and daily activity. *The Journals of Gerontology: Series B*, *75*(2), 377–388. https://doi.org/10.1093/geronb/gbz007

Galinsky, A. D., Turek, A., Agarwal, G., Anicich, E. M., Rucker, D. D., Bowles, H. R., Liberman, N., Levin, C., & Magee, J. C. (2024). Are many sex/gender differences really power differences? *PNAS Nexus*, *3*(2), Article pgae025. https://doi.org/10.1093/pnasnexus/pgae025

Haas, M., Mehl, M. R., Ballhausen, N., Zuber, S., Kliegel, M., & Hering, A. (2022). The sounds of memory: Extending the age–prospective memory paradox to everyday behavior and conversations. *The Journals of Gerontology: Series B*, *77*(4), 695–703. https://doi.org/10.1093/geronb/gbac012

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., Rentfrow, P. J., Campbell, A. T., & Gosling, S. D. (2020). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*, *119*(1), 204–228. https://doi.org/10.1037/pspp0000245

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big-Five Inventory—Version 4a and 54*. Berkeley Institute of Personality and Social Research, University of California.

Kaplan, D. M., Mehl, M. R., Pace, T. W., Negi, L. T., Silva, B. O. D., Lavelle, B. D., Sivilli, T., Williams, A., Comstock, T., Price, B., Medrano,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

V., Robbins, M. L., Cole, S. P., Craighead, W. E., & Raison, C. L. (2022). Implications of a "null" randomized controlled trial of mindfulness and compassion interventions in healthy adults. *Mindfulness*, *13*(5), 1197–1213. https://doi.org/10.1007/s12671-022-01861-3

Karpowitz, C. F., & Mendelberg, T. (2014). *The silent sex: Gender, deliberation, and Institutions*. Princeton University Press.

Kobayashi, Y., & Murakami, T. (2021, February 16). *Tokyo Olympic chief Mori to resign after sexist remarks*. The Mainichi. https://mainichi.jp/english/articles/20210211/p2a/00m/0na/004000c

Krackow, E., & Rudolph, K. D. (2008). Life stress and the accuracy of cognitive appraisals in depressed youth. *Journal of Clinical Child and Adolescent Psychology*, *37*(2), 376–385. https://doi.org/10.1080/15374410801955797

Kruschke, J. K. (2011). Introduction to special section on Bayesian data analysis. *Perspectives on Psychological Science*, *6*(3), 272–273. https://doi.org/10.1177/1745691611406926

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Landrine, H. (1985). Race × class stereotypes of women. *Sex Roles*, *13*(1–2), 65–75. https://doi.org/10.1007/BF00287461

Lazarević, L. B., Bjekić, J., Živanović, M., & Knežević, G. (2020). Ambulatory assessment of language use: Evidence on the temporal stability of electronically activated recorder and stream of consciousness data. *Behavior Research Methods*, *52*(5), 1817–1835. https://doi.org/10.3758/s13428-020-01361-z

Leaper, C., & Ayres, M. M. (2007). A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Personality and Social Psychology Review*, *11*(4), 328–363. https://doi.org/10.1177/1088868307302221

Liberman, M. (2006, August 6). *Language log. Language log: Sex-linked lexical budgets*. Retrieved September 28, 2021, from http://itre.cis.upenn.edu/myl/languagelog/archives/003420.html

Light, K. C., Grewen, K. M., & Amico, J. A. (2005). More frequent partner hugs and higher oxytocin levels are linked to lower blood pressure and heart rate in premenopausal women. *Biological Psychology*, *69*(1), 5–21. https://doi.org/10.1016/j.biopsycho.2004.11.002

Luo, M., Debelak, R., Schneider, G., Martin, M., & Demiray, B. (2021). With a little help from familiar interlocutors: Real-world language use in young and older adults. *Aging and Mental Health*, *25*(12), 2310–2319. https://doi.org/10.1080/13607863.2020.1822288

Macbeth, A., Bruni, M. R., De La Cruz, B., Erens, J. A., Atagi, N., Robbins, M. L., Chiarello, C., & Montag, J. L. (2022). Using the electronically activated recorder (EAR) to capture the day-to-day linguistic experiences of young adults. *Collabra: Psychology*, *8*(1), Article 36310. https://doi.org/10.1525/collabra.36310

Mangalindan, J. P. (2017, July 12). *LEAKED AUDIO: Uber's all-hands meeting had some uncomfortable moments*. Yahoo!finance. https://finance.yahoo.com/news/inside-ubers-hands-meeting-travis-194232221.html?soc_src=social-sh&soc_trk=tw

Mangelsdorf, S. N., Mehl, M. R., Qiu, J., & Alisic, E. (2019). How do mothers and fathers interact with their children after an injury? Exploring the role of parental acute stress, optimism, and self-efficacy. *Journal of Pediatric Psychology*, *44*(3), 311–322. https://doi.org/10.1093/jpepsy/jsy107

Manson, J. H. (2018). Life history strategy and everyday word use. *Evolutionary Psychological Science*, *4*(2), 111–123. https://doi.org/10.1007/s40806-017-0119-3

Manson, J. H., & Robbins, M. L. (2017). New evaluation of the electronically activated recorder (EAR): Obtrusiveness, compliance, and participant self-selection effects. *Frontiers in Psychology*, *8*, Article 658. https://doi.org/10.3389/fpsyg.2017.00658

Mascaro, J. S., Rentscher, K. E., Hackett, P. D., Lori, A., Darcher, A., Rilling, J. K., & Mehl, M. R. (2018). Preliminary evidence that androgen signaling is correlated with men's everyday language. *American Journal of Human Biology*, *30*(4), Article e23136. https://doi.org/10.1002/ajhb.23136

McCurry, J. (2021, February 13). Tokyo Olympics chief resigns over sexist comments. *The Guardian*. https://www.theguardian.com/sport/2021/feb/12/tokyo-olympics-chief-resigns-over-sexist-comments

Mehl, M. R. (2017). The electronically activated recorder (EAR): A method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science*, *26*(2), 184–190. https://doi.org/10.1177/0963721416680611

Mehl, M. R., & Holleran, S. E. (2007). An empirical analysis of the obtrusiveness of and participants' compliance with the electronically activated recorder (EAR). *European Journal of Psychological Assessment*, *23*(4), 248–257. https://doi.org/10.1027/1015-5759.23.4.248

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517–523. https://doi.org/10.3758/BF03195410

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*(5834), Article 82. https://doi.org/10.1126/science.1139940

Metcalf, C. A., & Dimidjian, S. (2020). In a mother's voice: Observing social–emotional aspects of postpartum daily life. *Journal of Family Psychology*, *34*(3), 269–278. https://doi.org/10.1037/fam0000587

Minor, K. S., Davis, B. J., Marggraf, M. P., Luther, L., & Robbins, M. L. (2018). Words matter: Implementing the electronically activated recorder in schizotypy. *Personality Disorders*, *9*(2), 133–143. https://doi.org/10.1037/per0000266

O'Hara, K. L., Grinberg, A. M., Tackman, A. M., Mehl, M. R., & Sbarra, D. A. (2020). Contact with an ex-partner is associated with psychological distress after marital separation. *Clinical Psychological Science*, *8*(3), 450–463. https://doi.org/10.1177/2167702620916454

Onnela, J.-P., Waber, B. N., Pentland, A., Schnorf, S., & Lazer, D. (2014). Using sociometers to quantify social interaction patterns. *Scientific Reports*, *4*(1), Article 5604. https://doi.org/10.1038/srep05604

Paruthi, S., Brooks, L. J., D'Ambrosio, C., Hall, W. A., Kotagal, S., Lloyd, R. M., Malow, B. A., Maski, K., Nichols, C., Quan, S. F., Rosen, C. L., Troester, M. M., & Wise, M. S. (2016). Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. *Journal of Clinical Sleep Medicine*, *12*(6), 785–786. https://doi.org/10.5664/jcsm.5866

Patterson, C. J., Sutfin, E. L., & Fulcher, M. (2004). Division of labor among lesbian and heterosexual parenting couples: Correlates of specialized versus shared patterns. *Journal of Adult Development*, *11*(3), 179–189. https://doi.org/10.1023/B:JADE.0000035626.90331.47

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Perrin, S., Meiser-Stedman, R., & Smith, P. (2005). The Children's Revised Impact of Event Scale (CRIES): Validity as a screening instrument for PTSD. *Behavioural and Cognitive Psychotherapy*, *33*(4), 487–498. https://doi.org/10.1017/S1352465805002419

Polsinelli, A. J., Moseley, S. A., Grilli, M. D., Glisky, E. L., & Mehl, M. R. (2020). Natural, everyday language use provides a window into the integrity of older adults' executive functioning. *The Journals of Gerontology: Series B*, *75*(9), e215–e220. https://doi.org/10.1093/geronb/gbaa055

Reddit. (2018). *haha get it cause women talk alot* [Meme]. https://www.reddit.com/r/ComedyCemetery/comments/8cstug/haha_get_it_cause_women_talk_alot/

Robbins, M. L., López, A. M., Weihs, K. L., & Mehl, M. R. (2014). Cancer conversations in context: Naturalistic observation of couples coping with breast cancer. *Journal of Family Psychology*, *28*(3), 380–390. https://doi.org/10.1037/a0036458

Robbins, M. L., Mehl, M. R., Holleran, S. E., & Kasle, S. (2011). Naturalistically observed sighing and depression in rheumatoid arthritis patients: A preliminary study. *Health Psychology*, *30*(1), 129–133. https://doi.org/10.1037/a0021558

Robbins, M. L., Spahr, C. M., & Karan, A. (2024). Re-evaluating the honing framework: Naturalistic observation of same- and different-gender couples' conversations. *Personal Relationships*, *31*(1), 156–174. https://doi.org/10.1111/pere.12533

Roos, Y., Krämer, M. D., Richter, D., Schoedel, R., & Wrzus, C. (2023). Does your smartphone "know" your social life? A methodological comparison of day reconstruction, experience sampling, and mobile sensing. *Advances in Methods and Practices in Psychological Science*, *6*(3), 1–12. https://doi.org/10.1177/25152459231178738

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*(2), 336–356. https://doi.org/10.1037/0033-295X.115.2.336

Schmitt, D. P. (2016, March 17). Sex differences in talkativeness? *Psychology Today*. https://www.psychologytoday.com/us/blog/sexual-personalities/201603/sex-differences-in-talkativeness

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schultheiss, O. C., Köllner, M. G., Busch, H., & Hofer, J. (2021). Evidence for a robust, estradiol-associated sex difference in narrative-writing fluency. *Neuropsychology*, *35*(3), 323–333. https://doi.org/10.1037/neu0000706

Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist*, *51*(10), 1065–1071. https://doi.org/10.1037/0003-066X.51.10.1065

Slatcher, R. B., & Robles, T. F. (2012). Preschoolers' everyday conflict at home and diurnal cortisol patterns. *Health Psychology*, *31*(6), 834–838. https://doi.org/10.1037/a0026774

Sun, J., & Vazire, S. (2019). Do people know what they're like in the moment? *Psychological Science*, *30*(3), 405–414. https://doi.org/10.1177/0956797618818476

Sung, Y. J., Schwander, K., Arnett, D. K., Kardia, S. L., Rankinen, T., Bouchard, C., Boerwinkle, E., Hunt, S. C., & Rao, D. C. (2014). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene–environment interactions. *Genetic Epidemiology*, *38*(4), 369–378. https://doi.org/10.1002/gepi.21800

Talbot, M. (2003). Gender stereotypes. Reproduction and challenge. In J. Holmes & M. Meyerhof (Eds.), *The handbook of language and gender* (pp. 468–486). Blackwell. https://doi.org/10.1002/9780470756942.ch20

Tamres, L. K., Janicki, D., & Helgeson, V. S. (2002). Sex differences in coping behavior: A meta-analytic review and an examination of relative coping. *Personality and Social Psychology Review*, *6*(1), 2–30. https://doi.org/10.1207/S15327957PSPR0601_1

Taylor, S. E., Gonzaga, G. C., Klein, L. C., Hu, P., Greendale, G. A., & Seeman, T. E. (2006). Relation of oxytocin to psychological stress responses and hypothalamic–pituitary-adrenocortical axis activity in older women. *Psychosomatic Medicine*, *68*(2), 238–245. https://doi.org/10.1097/01.psy.0000203242.95990.74

Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A. R., & Updegraff, J. A. (2000). Biobehavioral responses to stress in females: Tend-and-befriend, not fight-or-flight. *Psychological Review*, *107*(3), 411–429. https://doi.org/10.1037/0033-295X.107.3.411

Tornello, S. L. (2020). Division of labor among transgender and gender non-binary parents: Association with individual, couple, and children's behavioral outcomes. *Frontiers in Psychology*, *11*, Article 15. https://doi.org/10.3389/fpsyg.2020.00015

van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S. J., Valentine, K. D., van't Veer, A., & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*, Article 2625. https://doi.org/10.15626/MP.2020.2625

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology*, *95*(5), 1202–1216. https://doi.org/10.1037/a0013314

Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A., Kushida, C., Malhotra, R. K., Martin, J. L., Patel, S. R., Quan, S. F., & Tasali, E. (2015). Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Sleep*, *38*(6), 843–844. https://doi.org/10.5665/sleep.4716

Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A., Kushida, C., Malhotra, R. K., Martin, J. L., Patel, S. R., Quan, S. F., Tasali, E., Twery, M., Croft, J. B., Maher, E., Barrett, J. A., … the Consensus Conference Panel. (2015). Joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society on the recommended amount of sleep for a healthy adult: Methodology and discussion. *Sleep*, *38*(8), 1161–1183. https://doi.org/10.5665/sleep.4886

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214–227. https://doi.org/10.1177/2515245919848684

Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, *7*(1), 1–14. https://doi.org/10.1177/25152459231213802

Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2021). Centering categorical predictors in multilevel models: Best practices and interpretation. *Psychological Methods*, *28*(3), 613–630.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist*, *70*(1), 10–20. https://doi.org/10.1037/a0038208

Received November 30, 2021
Revision received September 10, 2024
Accepted September 17, 2024 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.