Title	Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses
Type	Article
URL	https://clok.uclan.ac.uk/id/eprint/45561/
DOI	https://doi.org/10.1177/25152459231162567
Date	2023
Citation	Coretta, Stefano, Casillas, Joseph V, Roessig, Simon, Franke, Michael, Ahn, Byron, Al-Hoorie, Ali H., Al-Tamimi, Jalal, Alotaibi, Najd E., AlShakhori, Mohammed K. et al (2023) Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses. Advances in Methods and Practices in Psychological Sciences, 6 (3). ISSN 2515-2459
Creators	Coretta, Stefano, Casillas, Joseph V, Roessig, Simon, Franke, Michael, Ahn, Byron, Al-Hoorie, Ali H., Al-Tamimi, Jalal, Alotaibi, Najd E., AlShakhori, Mohammed K., Altmiller, Ruth M., Arantes, Pablo, Athanasopoulou, Angeliki, Baese-Berk, Melissa M., Bailey, George, Baira A Sangma, Cheman, Beier, Eleonora J., Benavides, Gabriela M., Benker, Nicole, BensonMeyer, Emelia P., Benway, Nina R., Berry, Grant M., Bing, Liwen, Bjorndahl, Christina, Bolyanatz, Mariska, Braver, Aaron, Brown, Violet A., Brown, Alicia M., Brugos, Alejna, Buchanan, Erin M., Butlin, Tanna, Buxo-Lugo, Andres, Caillol, Coline, Cangemi, Francesco, Carignan, Christopher, Carraturo, Sita, Caudrelier, Tiphaine, Chodroff, Eleanor, Cohn, Michelle, Cronenberg, Johanna, Crouzet, Olivier, Dagar, Erica L., Dawson, Charlotte, Diantoro, Carissa A., Dokovova, Marie, Drake, Shiloh, Du, Fengting, Dubuis, Margaux, Dueme, Florent, Durward, Matthew, Egurtzegi, Ander, Elsherif, Mahmoud M., Esser, Janina, Ferragne, Emmanuel, Ferreira, Fernanda, Fink, Lauren K., Finley, Sara, Foster, Kurtis, Foulkes, Paul, Franzke, Rosa, Frazer-McKee, Gabriel, Fromont, Robert, Garcia, Christina, Geller, Jason, Grasso, Camille L., Greca, Pia, Grice, Martine, Grose-Hodge, Magdalena S., Gully, Amelia J., Halfacre, Caitlin, Hauser, Ivy, Hay, Jen, Haywood, Robert, Hellmuth, Sam, Hilger, Allison I., Holliday, Nicole, Hoogland, Damar, Huang, Yaqian, Hughes, Vincent, Icardo Isasa, Ane, Ilchovska, Zlatomira G., Jeon, Hae-Sung, Jones, Jacq, Junges, Magat N., Kaefer, Stephanie, Kaland, Constantijn, Kelley, Matthew C., Kelly, Niamh E., Kettig, Thomas, Khattab, Ghada, Koolen, Ruud, Krahmer, Emiel, Krajewska, Dorota, Krug, Andreas, Kumar, Abhilasha A., Lander, Anna, Lentz, Tomas O., Li, Wanyin, Li, Yanyu, Lialiou, Maria, Lima Jr., Ronaldo M., Lo, Justin J. H., Lopez Otero, Julio Cesar, Mackay, Bradley, MacLeod, Bethany, Mallard, Mel, McConnellogue, Carol-Ann Mary, Moroz, George, Murali, Mridhula, Nalborczyk, Ladislas, Nenadic, Filip, Nieder, Jessica, Nikolic, Dusan, Nogueira, Francisco G. S., Offerman, Heather M., Passoni, Elisa, Pelissier, Maud, Perry, Scott J., Pfiffner, Alexandra M., Proctor, Michael, Rhodes, Ryan, Rodriguez, Nicole, Roepke, Elizabeth, Roer, Jan P., Sbacco, Lucia, Scarborough, Rebecca, Schaeffler, Felix, Schleef, Erik, Schmitz, Dominic, Shiryaev, Alexander, Soskuthy, Marton, Spaniol, Malin, Stanley, Joseph, Strickler, Alyssa, Tavano, Alessandro, Tomaschek, Fabian, Tucker, Benjamin V., Turnbull, Rory, Ugwuanyi, Kingsley O., Urrestarazu-Porta, Inigo, van de Vijver, Ruben, Van Engen, Kristin J., van Miltenburg, Emiel, Wang, Bruce, Warner, Natasha, Wehrle, Simon, Westerbeek, Hans, Wiener, Seth, Winters, Stephen, Wong, Sidney G.-J., Wood, Anna, Wottawa, Jane, Xu, Chenzi, Zarate-Sanchez, German, Zellou, Georgia, Zhang, Cong, Zhu, Jian and Roettger, Timo B.

Registered Report

Advances in Methods and Practices in Psychological Science
July-September 2023, Vol. 6, No. 3, pp. 1–29

DOI: 10.1177/25152459231162567
www.psychologicalscience.org/AMPPS

Multidimensional Signals and Analytic Flexibility: Estimating Degrees of Freedom in Human-Speech Analyses

Stefano Coretta[1], Joseph V. Casillas[2], Simon Roessig[3], Michael Franke[4], Byron Ahn[5], Ali H. Al-Hoorie[6], Jalal Al-Tamimi[7], Najd E. Alotaibi[8], Mohammed K. AlShakhori[9], Ruth M. Altmiller[10], Pablo Arantes[11], Angeliki Athanasopoulou[12], Melissa M. Baese-Berk[13], George Bailey[14], Cheman Baira A Sangma[12], Eleonora J. Beier[15], Gabriela M. Benavides[9], Nicole Benker[16], Emelia P. BensonMeyer[17], Nina R. Benway[18], Grant M. Berry[19], Liwen Bing[20], Christina Bjorndahl[21], Mariška Bolyanatz[22], Aaron Braver[23], Violet A. Brown[10], Alicia M. Brown[24], Alejna Brugos[25], Erin M. Buchanan[26], Tanna Butlin[12], Andrés Buxó-Lugo[27], Coline Caillol[28], Francesco Cangemi[29], Christopher Carignan[30], Sita Carraturo[10], Tiphaine Caudrelier[31], Eleanor Chodroff[14], Michelle Cohn[32], Johanna Cronenberg[16], Olivier Crouzet[33], Erica L. Dagar[34], Charlotte Dawson[35], Carissa A. Diantoro[13], Marie Dokovova[36], Shiloh Drake[13], Fengting Du[37], Margaux Dubuis[38], Florent Duême[31], Matthew Durward[39], Ander Egurtzegi[40], Mahmoud M. Elsherif[41], Janina Esser[42], Emmanuel Ferragne[28], Fernanda Ferreira[15], Lauren K. Fink[43], Sara Finley[44], Kurtis Foster[13], Paul Foulkes[14], Rosa Franzke[16], Gabriel Frazer-McKee[45], Robert Fromont[46], Christina García[47], Jason Geller[48], Camille L. Grasso[49], Pia Greca[16], Martine Grice[29], Magdalena S. Grose-Hodge[20], Amelia J. Gully[14], Caitlin Halfacre[37], Ivy Hauser[34], Jen Hay[46], Robert Haywood[50], Sam Hellmuth[14], Allison I. Hilger[51], Nicole Holliday[52], Damar Hoogland[8], Yaqian Huang[53], Vincent Hughes[14], Ane Icardo Isasa[54], Zlatomira G. Ilchovska[55], Hae-Sung Jeon[56], Jacq Jones[39], Mágat N. Junges[57], Stephanie Kaefer[39], Constantijn Kaland[58], Matthew C. Kelley[59], Niamh E. Kelly[37], Thomas Kettig[14], Ghada Khattab[8], Ruud Koolen[60], Emiel Krahmer[60], Dorota Krajewska[61], Andreas Krug[8], Abhilasha A. Kumar[62], Anna Lander[63], Tomas O. Lentz[60], Wanyin Li[55], Yanyu Li[8], Maria Lialiou[64], Ronaldo M. Lima, Jr.[65], Justin J. H. Lo[30],

Corresponding Author:
Timo B. Roettger, Department of Linguistics and Scandinavian Studies, University of Oslo, Norway
Email: timo.b.roettger@iln.uio.no

Julio Cesar Lopez Otero[66], Bradley Mackay[67], Bethany MacLeod[68], Mel Mallard[10], Carol-Ann Mary McConnellogue[69], George Moroz[63], Mridhula Murali[70], Ladislas Nalborczyk[49], Filip Nenadić[71], Jessica Nieder[72], Dušan Nikolić[12], Francisco G. S. Nogueira[73], Heather M. Offerman[74], Elisa Passoni[75], Maud Pélissier[28], Scott J. Perry[76], Alexandra M. Pfiffner[77], Michael Proctor[78], Ryan Rhodes[79], Nicole Rodríguez[2], Elizabeth Roepke[80], Jan P. Röer[81], Lucia Sbacco[8], Rebecca Scarborough[82], Felix Schaeffler[83], Erik Schleef[67], Dominic Schmitz[84], Alexander Shiryaev[63], Márton Sóskuthy[85], Malin Spaniol[86], Joseph A. Stanley[87], Alyssa Strickler[82], Alessandro Tavano[88], Fabian Tomaschek[89], Benjamin V. Tucker[76], Rory Turnbull[37], Kingsley O. Ugwuanyi[90], Iñigo Urrestarazu-Porta[40], Ruben van de Vijver[72], Kristin J. Van Engen[10], Emiel van Miltenburg[60], Bruce Xiao Wang[91], Natasha Warner[9], Simon Wehrle[29], Hans Westerbeek[92], Seth Wiener[93], Stephen Winters[12], Sidney G.-J. Wong[94], Anna Wood[13], Jane Wottawa[95], Chenzi Xu[96], Germán Zárate-Sández[97], Georgia Zellou[32], Cong Zhang[8], Jian Zhu[98], and Timo B. Roettger[99]

[1]Department of Linguistics and English Language, University of Edinburgh, Edinburgh, United Kingdom; [2]Department of Spanish and Portuguese, Rutgers University, New Brunswick, New Jersey; [3]Department of Linguistics, Cornell University, Ithaca, New York; [4]Department of General and Computational Linguistics, University of Tübingen, Tübingen, Germany; [5]Program in Linguistics, Princeton University, Princeton, New Jersey; [6]Jubail English Language and Preparatory Year Institute, Royal Commission for Jubail and Yanbu, Jubail, Saudi Arabia; [7]Université Paris Cité, Laboratoire de Linguistique Formelle, CNRS, Paris, France; [8]School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [9]Department of Linguistics, University of Arizona, Tucson, Arizona; [10]Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, Missouri; [11]Departamento de Letras, Universidade Federal de São Carlos, São Carlos, Brazil; [12]School of Languages, Linguistics, Literatures and Cultures, University of Calgary, Calgary, Alberta, Canada; [13]Department of Linguistics, University of Oregon, Eugene, Oregon; [14]Department of Language and Linguistic Science, University of York, York, United Kingdom; [15]Department of Psychology, University of California, Davis, Davis, California; [16]Institute of Phonetics and Speech Processing, University of Munich, Munich, Germany; [17]Department of Linguistics, University of Pennsylvania, Philadelphia, Pennsylvania; [18]Department of Communication Sciences and Disorders, Syracuse University, Syracuse, New York; [19]Department of Spanish, Villanova University, Villanova, Pennsylvania; [20]Department of English Language and Linguistics, University of Birmingham, Birmingham, United Kingdom; [21]Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania; [22]Department of Spanish and French Studies, Occidental College, Los Angeles, California; [23]Department of English, Texas Tech University, Lubbock, Texas; [24]Department of Spanish and Portuguese, University of Arizona, Tucson, Arizona; [25]Division of Mathematics and Computer Science, Simmons University, Boston, Massachusetts; [26]Analytics, Harrisburg University of Science and Technology, Harrisburg, Pennsylvania; [27]Department of Psychology, University at Buffalo, Buffalo, New York; [28]CLILLAC-ARP, Université Paris Cité, Paris, France; [29]IfL-Phonetik, University of Cologne, Cologne, Germany; [30]Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom; [31]Basque Center on Cognition Brain and Language, Donostia, Spain; [32]Department of Linguistics, University of California, Davis, Davis, California; [33]Laboratoire de Linguistique de Nantes (UMR6310), Nantes Université, CNRS, Nantes, France; [34]Department of Linguistics and TESOL, University of Texas at Arlington, Arlington, Texas; [35]School of Psychology, Newcastle University, Newcastle upon Tyne, United Kingdom; [36]School of Psychological Sciences and Health, University of Strathclyde, Glasgow, United Kingdom; [37]School of English Literature, Language and Linguistics, Newcastle University, Newcastle upon Tyne, United Kingdom; [38]Department of Comparative Language Science, University of Zurich, Zurich, Swizerland; [39]Department of Linguistics, University of Canterbury, Christchurch, New Zealand; [40]IKER (UMR 5478), CNRS, Bayonne, France; [41]Department of Neuroscience, Psychology and Behaviour, University of Leicester, Leicester, United Kingdom; [42]Statistics Group, Association for Diversity in Linguistics, Cologne, Germany; [43]Department of Music, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany; [44]Department of Psychology, Pacific Lutheran University, Parkland, Washington; [45]Department of Languages, Linguistics, and Translation, Université Laval, Quebec City, Quebec, Canada; [46]New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, Christchurch, New Zealand; [47]Department of Languages, Literatures, and Cultures, Saint Louis University, St. Louis, Missouri; [48]Department of Psychology, Princeton University, Princeton, New Jersey; [49]LPC, Aix-Marseille

Université, CNRS, France; [50]Ao Tawhiti Unlimited Discovery, Christchurch, New Zealand; [51]Department of Speech, Language, and Hearing Sciences, University of Colorado Boulder, Boulder, Colorado; [52]Department of Linguistics and Cognitive Science, Pomona College, Claremont, California; [53]University of California, Los Angeles, Los Angeles, California; [54]Department of Modern and Classical Languages and Literatures, California State University, Northridge, California; [55]School of Psychology, University of Birmingham, Birmingham, United Kingdom; [56]School of Humanities, Language and Global Studies, University of Central Lancashire, Preston, United Kingdom; [57]Programa de Pós-Graduação em Letras, Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [58]Institute of Linguistics, University of Cologne, Cologne, Germany; [59]Department of Linguistics, University of Washington, Seattle, Washington; [60]Department of Communication and Cognition, Tilburg University, Tilburg, the Netherlands; [61]Department of Linguistics and Basque Studies, University of the Basque Country UPV/EHU, Vitoria-Gasteiz, Spain; [62]Department of Psychology, Bowdoin College, Brunswick, Maine; [63]Linguistic Convergence Laboratory, HSE University, Moscow, Russia; [64]Institute of German Language I Linguistics, University of Cologne, Cologne, Germany; [65]Department of English Language Studies, Federal University of Ceará, Fortaleza, Brazil; [66]Department of Hispanic Studies, University of Houston, Houston, Texas; [67]Department of English and American Studies, University of Salzburg, Salzburg, Austria; [68]School of Linguistics and Language Studies, Carleton University, Ottawa, Ontario, Canada; [69]Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; [70]Speech and Language Therapy, University of Strathclyde, Glasgow, United Kingdom; [71]Department of Psychology, Faculty of Media and Communications, Singidunum University, Belgrade, Serbia; [72]Institute of Linguistics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany; [73]Graduate Program in Linguistics, Federal University of Ceará, Fortaleza, Brazil; [74]Department of World Languages, Literatures and Cultures, University of Arkansas, Fayetteville, Arkansas; [75]Deparment of Linguistics, Queen Mary University of London, London, United Kingdom; [76]Department of Linguistics, University of Alberta, Alberta, Canada; [77]Department of Linguistics, University of California, Berkeley, Berkeley, California; [78]Department of Linguistics, Macquarie University, Sydney, Australia; [79]Center for Cognitive Science, Rutgers University, Piscataway, New Jersey; [80]Department of Speech, Language, and Hearing Sciences, Saint Louis University, St. Louis, Missouri; [81]Department of Psychology, Witten/Herdecke University, Witten, Germany; [82]Department of Linguistics, University of Colorado Boulder, Boulder, Colorado; [83]Clinical Audiology, Speech and Language Research Centre, Queen Margaret University Edinburgh, Edinburgh, United Kingdom; [84]Department of English and American Studies, Heinrich Heine University Düsseldorf, Düsseldorf, Germany; [85]Department of Linguistics, University of British Columbia, Vancouver, British Columbia, Canada; [86]Department of Psychiatry and Psychotherapy, University Hospital Cologne, Cologne, Germany; [87]Department of Linguistics, Brigham Young University, Provo, Utah; [88]Department of Neuroscience, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany; [89]Department of General Linguistics, University of Tübingen, Tübingen, Germany; [90]Department of English and Literary Studies, University of Nigeria, Nsukka, Nigeria; [91]Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong SAR, China; [92]Department of Languages, Literature and Communication, Utrecht University, Utrecht, the Netherlands; [93]Department of Modern Languages, Carnegie Mellon University, Pittsburgh, Pennsylvania; [94]Geospatial Research Institute, University of Canterbury, Christchurch, New Zealand; [95]Département de Lettres Modernes, LIUM, LST, Le Mans Université, Le Mans, France; [96]Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, United Kingdom; [97]Department of Spanish, Western Michigan University, Kalamazoo, Michigan; [98]School of Information, University of Michigan, Ann Arbor, Michigan; and [99]Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo, Norway

Abstract

Recent empirical studies have highlighted the large degree of analytic flexibility in data analysis that can lead to substantially different conclusions based on the same data set. Thus, researchers have expressed their concerns that these researcher degrees of freedom might facilitate bias and can lead to claims that do not stand the test of time. Even greater flexibility is to be expected in fields in which the primary data lend themselves to a variety of possible operationalizations. The multidimensional, temporally extended nature of speech constitutes an ideal testing ground for assessing the variability in analytic approaches, which derives not only from aspects of statistical modeling but also from decisions regarding the quantification of the measured behavior. In this study, we gave the same speech-production data set to 46 teams of researchers and asked them to answer the same research question, resulting in substantial variability in reported effect sizes and their interpretation. Using Bayesian meta-analytic tools, we further found little to no evidence that the observed variability can be explained by analysts' prior beliefs, expertise, or the perceived quality of their analyses. In light of this idiosyncratic variability, we recommend that researchers more transparently share details of their analysis, strengthen the link between theoretical construct and quantitative system, and calibrate their (un)certainty in their conclusions.

Keywords

crowdsourcing science, data analysis, scientific transparency, speech, acoustic analysis

Received 5/13/21; Revision accepted 2/15/23

To effectively accumulate knowledge, science needs to (a) produce data that can be replicated using the original methods and (b) arrive at robust conclusions substantiated by such data. In recent coordinated efforts to replicate published findings, scientific disciplines have uncovered surprisingly low success rates (e.g., Camerer et al., 2018; Open Science Collaboration, 2015), leading to what is now referred to as the replication crisis. Beyond the difficulties of replicating scientific findings, a growing body of evidence suggests that researchers' conclusions often vary even when they have access to the same data. The latter situation has been referred to as the inference crisis (Rotello et al., 2015; Starns et al., 2019) and is, among other things, rooted in the inherent flexibility of data analysis (often referred to as researcher degrees of freedom; Gelman & Loken, 2014; Simmons et al., 2011). Data analysis involves many different steps, such as inspecting, organizing, transforming, and modeling data, to name a few. Along the way, different methodological and analytic choices need to be made, all of which may influence the final interpretation of the data.

These researcher degrees of freedom are both a blessing and a curse. They are a blessing because they afford us the opportunity to look at nature from different angles, which, in turn, allows us to make important discoveries and generate new hypotheses (e.g., Box, 1976; De Groot, 2014; Tukey, 1977). They are a curse because idiosyncratic choices can lead to categorically different interpretations that eventually find their way into the publication record, where they are taken for granted (Simmons et al., 2011). Recent projects have shown that the variability between different data analysts is vast and can lead independent researchers to draw different conclusions from the same data set (e.g., Botvinik-Nezer et al., 2020; Silberzahn et al., 2018; Starns et al., 2019). These studies, however, might still underestimate the extent to which analysts vary because data analysis is not restricted to the statistical analysis of ready-made numeric data. These data can in fact be the result of complex measurement processes that translate a phenomenon, such as human behavior, into numbers. This is particularly true for fields that draw conclusions about human behavior and cognition from multidimensional data such as audio or video data. In fields working on speech production, for example, researchers need to make numerous decisions about what to measure and how to measure it (i.e., how to operationalize the phenomenon under investigation). Given the temporal extension of the acoustic signal and its complex structural composition, this is not trivial.

In this article, we investigate the impact of analytic choices on research results when many analyst teams examine the same speech-production data set, a process that involves both decisions regarding the operationalization of linguistically relevant constructs and decisions regarding statistical analysis. Specifically, we discuss the degree of variability in research results obtained by 46 teams who had to choose the operationalization and statistical procedures to answer the same research question on the basis of the same set of raw data (here, speech recordings). Our study seeks to (a) conceptually replicate previous many-analysts projects by probing the effects of different statistical analyses and by assessing the generalizability of published findings to other disciplines (here, the speech sciences) and (b) extend the scope of inquiry to include flexibility in the operationalization of complex human behavior (here, speech). This is an important addition in that the increased number of "forking paths" in the "garden of analytic choices"—derived from the many decisions involved in quantification—might reveal a higher degree of variability across analysts than previously observed, thus giving us a more realistic estimate of variability.

Researcher Degrees of Freedom

Data analysis comes with many decisions, such as how to measure a given phenomenon or behavior, which data to submit to statistical modeling and which to exclude in the final analysis, or what inferential decision-making procedure to apply. This can be problematic because humans show cognitive biases that can lead to erroneous inferences (Tversky & Kahneman, 1974). For example, humans see coherent patterns in randomness (Brugger, 2001), convince themselves of the validity of prior expectations ("I knew it"; Nickerson, 1998), and perceive events as being plausible in hindsight ("I knew it all along"; Fischhoff, 1975). In conjunction with an academic incentive system that rewards certain discovery processes more than others (Koole & Lakens, 2012; Sterling, 1959), we often find ourselves exploring many possible analytic pipelines but reporting only a selected few.

This issue is particularly amplified in fields in which the raw data lend themselves to many possible ways of being measured (Roettger, 2019). Combined with a wide variety of methodological and theoretical traditions as well as varying levels of quantitative training across subfields, the inherent flexibility of data analysis might lead to a vast plurality of analytic approaches that can lead to different scientific conclusions (Roettger et al., 2019). Analytic flexibility has been widely discussed from a conceptual point of view (Nosek & Lakens, 2014; Simmons et al., 2011; Wagenmakers et al., 2012) and in regard to its application in individual scientific fields (e.g., Charles et al., 2019; Roettger, 2019; Wicherts et al., 2016). This notwithstanding, there are still many unknowns regarding the extent of analytic plurality in practice.

Consequently, a substantial body of published articles likely present overconfident interpretations of data and statistical results based on idiosyncratic analytic strategies

(e.g., Gelman & Loken, 2014; Simmons et al., 2011). These interpretations, and the conclusions that derive from them, are thus associated with an unknown degree of uncertainty (dependent on the strength of evidence provided) and with an unknown degree of generalizability (dependent on the chosen analysis). Moreover, the same data could lead to very different conclusions depending on the analytic path taken by the researcher. However, instead of being critically evaluated, scientific results often remain unchallenged in the publication record. Despite recent efforts to improve transparency and reproducibility (e.g., Klein et al., 2018; Miguel et al., 2014) and the advent of freely available and accessible infrastructures, such as those provided by OSF, critical reanalyses of published analytic strategies are still uncommon because data sharing remains rare (Wicherts et al., 2006).

Crowdsourcing Alternative Analyses

Recent collaborative attempts have started to shed light on how different analysts tackle the same data set and have revealed a large amount of variability. In a pioneering collaborative effort, Silberzahn et al. (2018) let 29 independent analysis teams address the same research hypothesis: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players. The analytic approaches, and thus the results, varied widely between teams. Twenty teams (69%) found support for the hypothesis, and nine (31%) did not. Of the 29 analytic strategies, there were 21 unique combinations of covariates. Importantly, the observed variability was neither predicted by the teams' preconceptions about the phenomenon under investigation nor by peer ratings of the quality of their analyses. The authors' results suggest that analytic plurality may be an inevitable by-product of the scientific process and not necessarily driven by different levels of expertise or bias.

Several other recent studies have corroborated this analytic flexibility across different disciplines. Dutilh et al. (2019) and Starns et al. (2019) investigated analysts' choices when inferring theoretical constructs based on the same data set using computational models. Both studies revealed vastly different modeling strategies, even though scientific conclusions were similar across analysis teams (for analytic flexibility in neuroimaging data and ecology, respectively, see also Botvinik-Nezer et al., 2020; Parker et al., 2020). Bastiaansen et al. (2020) crowdsourced clinical recommendations based on analyses of an individual patient. Their results suggest that analysts differed substantially regarding decisions related to both the statistical analysis of the data and the theoretical rationale behind interpreting the statistical results.

Building on the many-analysts approach, Landy et al. (2020) asked 15 research teams to independently design studies to answer five different research questions related to moral judgments. Again, they found vast heterogeneity across researchers' conclusions. The observed variation was not predicted by the researchers' expertise but seem to vary for the five different research questions that might exhibit different degrees of theoretical underspecification. This is in line with Auspurg and Brüderl (2021), who reanalyzed the red-card study mentioned above. The authors argued that some of the observed heterogeneity across analysts in Silberzahn et al. (2018) might have been driven by flexibility in statistically interpreting the research question.

Although these studies attested to a large degree of analytic flexibility with possibly impactful consequences, they focused on analytic decisions related to the study design, the statistical analysis, or the architecture of computational models. In these studies the data sets were fixed, and neither data collection nor measurement could be changed. Thus, the estimates of variability found in the literature might reflect a lower bound only, ignoring large parts of the forking paths related to measurement. However, in many fields the primary raw data are complex signals for which theoretical constructs need to be operationalized relative to a theoretically motivated research question. This is especially true in the social sciences, in which the phenomenon under investigation corresponds to both observable and unobservable human behavior.

Decisions about how to measure theoretical constructs related to human behavior and cognition might interact with downstream decisions about statistical modeling and vice versa. For instance, Flake and Fried (2020) discussed the cascading impact that different practices can have on psychometric research. The authors highlighted, for example, the following degrees of freedom in the choice and development of measures: definition of the theoretical construct, justification of the selected measure, description of the measure and how it maps onto the construct, response coding and related transformations, as well as post hoc modifications to the chosen measure. Taken together, these aspects alone dramatically increase the combinations of possible analytic choices and hence flexibility in research outcomes.

In those disciplines concerned with communication, human behavior often corresponds to multidimensional visual and/or acoustic signals. The complex nature of these data exponentiates the number of possible analytic approaches, thus further increasing analytic flexibility. To estimate this increased flexibility, the current study looks at experimentally elicited speech production data.

Operationalizing Speech

Research on speech lies at the intersection of the cognitive sciences, informing psychological models of language, categorization, and memory; guiding methods for

the diagnosis and treatment of speech disorders; and facilitating advancement in automatic speech recognition and speech synthesis. One major challenge in the speech sciences is the mapping between communicative intentions (the unobserved behavior) and their physical manifestation (the observed behavior).

Speech signals are complex because they are characterized by structurally different acoustic parameters distributed throughout different temporal domains. Thus, choosing how to assess a communicative intention of interest is an important analytic step. Take, for example, the sentence in (1):

(1) "I can't bear another meeting on Zoom."

Depending on the speaker's intention, this sentence can be said in different ways. For instance, if the speaker is exhausted by all their meetings, they might acoustically highlight the word "another" or "meeting" to contrast it with more pleasant activities. If, on the other hand, the speaker is just tired of video conferences, as opposed to say face-to-face meetings, they might acoustically highlight the word "Zoom."

If we decide to compare the speech signal associated with these two intentions, how can we quantify the difference between them? In other words, given their physical manifestation (speech), what do we measure and how do we measure it? Because of the continuous and transient nature of speech, identifying speech parameters and temporal domains within which to measure those parameters becomes a nontrivial task. Utterances stretch over several thousand milliseconds and contain different levels of linguistically relevant units such as phrases, words, syllables, and individual sounds. The researcher is thus confronted with a considerable number of parameters and combinations thereof to choose from.

From a phonetic viewpoint, linguistically relevant units are inherently multidimensional and dynamic: They consist of clusters of parameters that are modulated over time. The acoustic parameters of units are usually asynchronous; that is, they appear at different time points in the unfolding signal and overlap with parameters of other units (e.g., Jongman et al., 2000; Lisker, 1986; Summerfield, 1981; Winter, 2014). A classic example is the distinction between voiced and voiceless stops in English (i.e., /b/ and /p/ in "bear" vs. "pear"). This contrast is manifested by many acoustic features that can differ depending on several factors, such as the position of the consonant in the word and context of surrounding sounds (Lisker, 1977). Furthermore, correlates of the contrast can even be found away from the consonant in temporally distant speech units. For example, the initial /l/ of the English words "led" and "let" is affected by the voicing of the final consonant (/d, t/; Hawkins & Nguyen, 2004).

The multiplicity of phonetic measurements grows exponentially if we look at larger temporal domains, as is the case with suprasegmental aspects of speech. For example, studies investigating acoustic correlates of word stress (e.g., the difference between "ínsight" and "incíte") use a wide variety of measurements, including temporal characteristics (duration of certain segments or subsegmental intervals), spectral characteristics (intensity, formants, and spectral tilt), and measurements related to fundamental frequency (f_0; e.g., Gordon & Roettger, 2017). Moving on to the expression of higher level communicative functions, such as information structure and discourse pragmatics, relevant acoustic cues can be distributed throughout even larger domains, such as phrases and whole utterances (e.g., Ladd, 2008). Differences in the position, shape, and alignment of f_0 modulations over multiple locations within a sentence are correlated with differences in discourse functions (e.g., Niebuhr et al., 2011). The latter can also be expressed by global versus local pitch modulations (Van Heuven et al., 2002), as well as acoustic information within the temporal or spectral domain (e.g., Van Heuven & Van Zanten, 2005). Extralinguistic information, such as the speaker's intentions, levels of emotional arousal, or social identity, are also conveyed by broad domain parameters, such as voice quality, rhythm, and pitch (Foulkes & Docherty, 2006; Ogden, 2004; White et al., 2009).

In short, when testing hypotheses on speakers' intentions using speech production data, researchers are faced with many choices and possibilities. The larger the functional domain (e.g., segments vs. words vs. utterances), the higher the number of conceivable operationalizations. For example, several decisions have to be made when comparing the two realizations of the sentence in (1), one of which is intended to signal emphasis on "another" and one of which emphasizes "Zoom":

(2a) I can't bear *another* meeting on Zoom.

(2b) I can't bear another meeting on *Zoom*.

Do we compare only the word "another" in (2a) and (2b) or also the word "Zoom"? Do we measure utterance-wide acoustic profiles, whole words, or just stressed syllables? Do we average across the chosen time domain or do we measure a specific point in time? Do we measure f_0, intensity, or something else (Stevens, 2000)?

When looking at phrase-level temporal domains, the number of possible alternative analytic pipelines increases substantially. Figure 1a shows a typical example of a decision tree with which speech researchers are often confronted. Each of the four analytic decisions in

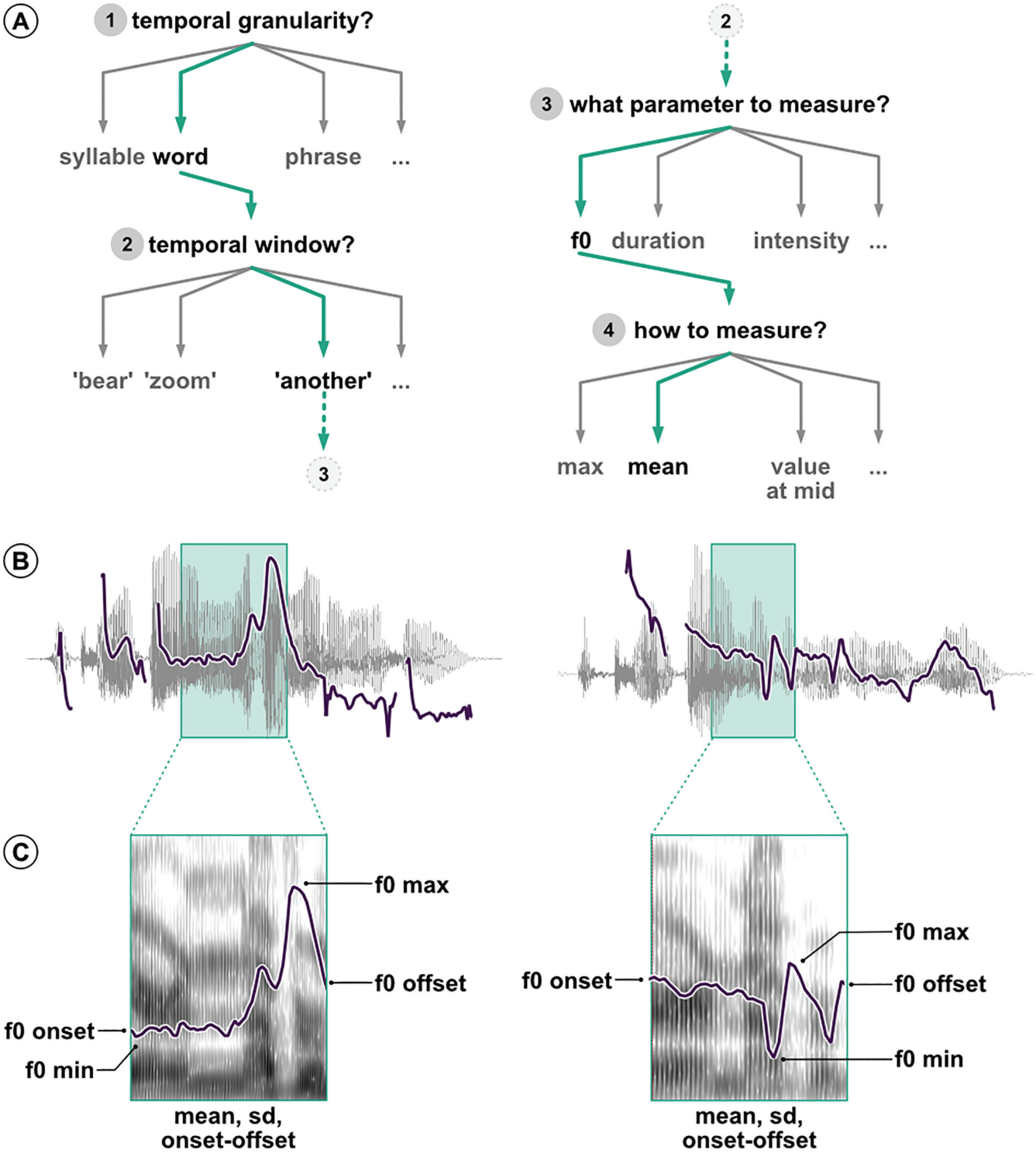

Fig. 1. Illustration of the analytic flexibility associated with acoustic analyses. (a) Example of multiple possible and justifiable decisions when comparing two utterances. (b) Waveform and f_0 track of the utterances "I can't bear *another* meeting on Zoom" and "I can't bear another meeting on *Zoom*." The green boxes mark the word "another" in both sentences. (c) Spectrogram and f_0 track of the word "another," exemplifying possible operationalizations of differences in f_0.

the example have different possible options. Here only one particular path has been taken. A different one would likely produce different results and might lead to different conclusions. Once we have decided to compare f_0 of the word "another" across the two utterances, there are still many choices to be made, all of which need to

be justified. As Figures 1b and 1c illustrate, we could measure f_0 at specific points in time such as the onset of the temporal window, the offset, or the midpoint. We could also measure the value or time of the f_0 minimum or maximum. We could summarize f_0 across the entire window and extract the mean, median, or standard deviation of f_0, all of which have been used to analyze speech data in previous work (see Gordon & Roettger, 2017). But the journey in the garden of analytic paths goes on. Other important operationalization steps could involve filtering the audio signal, smoothing the extracted f_0 track, removing values that substantially deviate from surrounding values or expectations, either manually or automatically, and so on.

These decisions are intended to be made prior to any statistical analysis but are at times revised a posteriori in light of unforeseen or surprising outcomes (i.e., after data collection and/or preliminary analyses). This multitude of possible decisions is multiplied by those researcher degrees of freedom related to statistical analysis (e.g., Wicherts et al., 2016).

In sum, speech data are made of complex physical signals that generate an as-of-yet unappreciated amount of analytic flexibility in the choice of measures and operationalizations. This article probes this garden of forking paths in the analysis of speech. To assess the variability in data-analysis pipelines, including both operationalization and statistical analysis, across independent researchers, we provide analytic teams with an experimentally elicited speech-production data set. The data set derives from the unpublished research project "Prosodic Encoding of Redundant Referring Expressions," which set out to investigate whether speakers acoustically modify utterances to signal unexpected referring expressions.[1] In the following section we introduce the research question and the experimental procedure of this project and describe the resulting data set as used in the current study.

The Data Set: Acoustic Properties of Atypical Modifiers

Referring is one of the most basic and prevalent uses of language and one of the most widely researched areas in language science. When trying to refer to a banana, what does a speaker say and how do they say it in a given context? The context within which an entity occurs (i.e., with other nonfruits, other fruits, or other bananas) plays a large part in determining the choice of referring expressions. Speakers generally aim to be as informative as possible to uniquely establish reference to the intended object, but they are also resource-efficient in that they avoid redundancy (Grice, 1975). Thus, one would expect the use of a modifier, for example, only if it is necessary for disambiguation. For instance, one might use the adjective "yellow" to describe a banana in a situation in which there are both a yellow and a less ripe green banana available, but not when there is only one banana.

Despite the coherent idea that speakers are both rational and efficient, there is much evidence that speakers are often overinformative. Speakers use referring expressions that are more specific than strictly necessary for the unambiguous identification of the intended referent (Rubio-Fernández, 2016; Sedivy, 2003), which has been argued to facilitate object identification and make communication between speakers and listeners more efficient (Arts et al., 2011; Paraboni et al., 2007; Rubio-Fernández, 2016). Recent findings suggest that the utility of referring expressions depends on how useful they are for a listener (compared with other referring expressions) to identify a target object. For example, Degen et al. (2020) showed that modifiers that are less typical for a given referent (e.g., a blue banana) are more likely to be used in an overinformative scenario (e.g., when there is just one banana; see also Westerbeek et al., 2015). This account, however, has mainly focused on content selection (Gatt et al., 2011), that is, what words to use.

Even when morphosyntactically identical expressions are involved, speakers can modulate utterances via acoustic properties such as temporal and spectral modifications (e.g., Ladd, 2008). Most prominently, languages can use intonation to signal discourse relationships between referents. Intonation marks discourse-relevant referents for being new or given information to guide the listeners' interpretation of incoming messages. Beyond structuring information relative to the discourse, a few studies have suggested that speakers might use intonation to signal atypical lexical combinations (e.g., Dimitrova et al., 2008, 2009). Referential expressions such as "blue banana" were produced with greater prosodic prominence than more typical referents such as "yellow banana." These results are in line with the idea of resource-efficient, rational language users who modulate their speech to facilitate listeners' comprehension. However, the above studies are based on a small sample size (10 participants) and on potentially anticonservative statistical analyses, leaving reason to doubt the generalizability of the studies' conclusions.

To further illuminate the question of whether speakers modify speech to signal atypical referents and overcome some of the limitations of previous work, 30 native German speakers were recorded in a production study while interacting with a confederate (one of the experimenters) in a referential game, following experimental procedures typical of the field. The participants had to verbally instruct the confederate to select a specified target object out of four objects presented on a screen.

The subject and confederate were seated at opposite sides of a table, each facing one of two computer screens. The participant and the experimenter could not see each other or each others' screens. Figure 2 shows the experimental-procedure time line. After a familiarization phase, the subject first saw four colored objects in the top-left, top-right, bottom-left, and bottom-right corners of the screen. One of the objects served as the target, another served as the competitor, and the remaining two objects served as distractors. Objects were referred to using noun phrases consisting of an adjective modifier denoting color and a modified object (e.g., *gelbe Zitrone*, "yellow lemon"; *rote Gurke*, "red cucumber"; *rote Socken*, "red socks").

In the center of the screen, a black cube was displayed that could be moved by the experimenter. The participants read a sentence prompt out loud (*Du sollst den Würfel auf der COLOR OBJECT ablegen*; "You have to put the cube on top of the COLOR OBJECT") to instruct the experimenter to drag the cube on top of one of the four depicted objects (the *competitor*) using the mouse. After the experimenter had moved the cube as instructed, the subject would read another sentence prompt (*Und jetzt sollst du den Würfel auf der COLOR OBJECT ablegen*; "And now, you have to put the cube on top of the COLOR OBJECT"), instructing the experimenter to move the cube on top of a different object (the *target*). The second utterance in the trial was the critical trial for analysis.

The two sentence prompts were used to create a focus contrast between the competitor and the target object. Focused units denote the set of all (contextually relevant) alternatives (e.g., Rooth, 1992). Concretely, a focus contrast marks one or more elements in a sentence as prominent by different linguistic means depending on the language (Burdin et al., 2015; Mati'c & Wedgwood, 2013). For instance, if the competitor and target objects differ but their color does not (e.g., yellow banana vs. yellow tomato), the noun is said to be in focus (noun-focus, or NF, condition). If the objects are the same but differ in color (e.g., yellow banana vs. blue banana), the color adjective is in focus (adjective-focus, or AF, condition). If both the color and the object differ (e.g., yellow banana vs. blue tomato), then the whole noun phrase is in focus (adjective/noun-focus, or ANF, condition). The NF condition constituted the experimentally relevant condition, whereas the AF and ANF conditions acted as fillers. Crucially, the color-object combinations in the NF condition were manipulated with respect to their typicality. The combinations were either typical (e.g., orange mandarin), medium typical (e.g., green tomato), or atypical (e.g., yellow cherry), as established by a norming study that was conducted prior to the production experiment just described.[2] Each subject produced 15 critical trials (NF condition). Each trial was repeated twice, yielding a total of 30 trials per participant and a grand total of 900 ($15 \times 2 \times 30$ participants) spoken utterances.

For the current study, 46 analysis teams received access to the entire data set generated by the production study. The data set is constituted by audio recordings and annotation files in a format that is typical for the field. The teams were instructed to answer the following research question using the provided data set: Do speakers acoustically modify utterances to signal atypical word combinations?

Method

As outlined in the Operationalizing Speech section, researchers are faced with a large number of analytic choices when analyzing a multidimensional signal such as speech. Analysts must identify and operationalize relevant measurements, as well as the temporal domain(s) from which these measurements are to be taken, and then possibly transform these measurements before submitting them to statistical models, which must be chosen alongside inferential criteria. The complexity of speech data constitutes the ideal testing ground for assessing the upper bound of analytic flexibility that social scientists might face across disciplines. We used a meta-analytic approach to assess (a) the variability of the reported effects and (b) how **analytic and researcher-related predictors** affect the final results (bold terms are defined in the Glossary in the Appendix).

In this study, we followed the procedures proposed by Parker et al. (2020) and Aczel et al. (2021). The project comprised the following five phases:

1. Recruitment: We recruited independent groups of researchers to analyze the data and review others' data analyses.
2. Team analysis: We gave researchers access to the speech corpus and let them analyze the data as they saw fit.
3. Review: We asked reviewers to generate peer-review ratings of the analyses based on methods (not results).
4. Meta-analysis: We evaluated variability among the different analyses and how different predictors affected the outcomes.
5. Write-up: We collaboratively produced the final manuscript.

We initially estimated that this process, from the time of an in-principle acceptance of the Stage 1 registered report to the end of Phase 5, would take 9 months. Phase

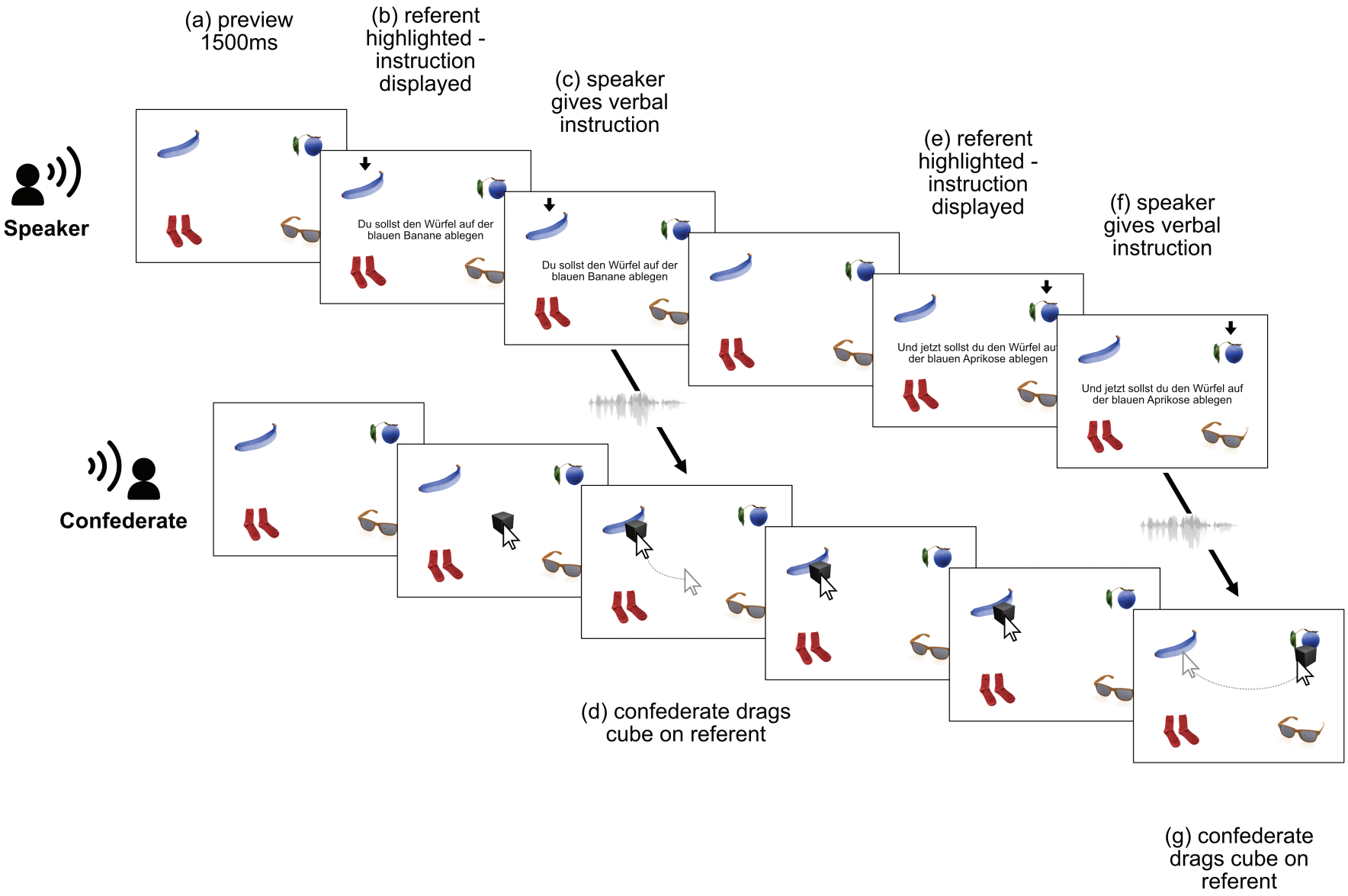

Fig. 2. Experimental procedure. The top row illustrates the trial sequence for the speaker (participant), and the bottom row illustrates the trial sequence for the confederate. After a preview of 1,500 ms (a), the speaker sees an arrow indicating one of the referents (b). Reading the orthographic instructions out loud, the speaker gives the confederate verbal instructions onto which referent they should drag the cube (c). The confederate, in turn, drags the black cube onto the target referent (d). Both the arrow and the orthographic instruction disappear from the speaker's screen, and a new referent is indicated by an arrow on the same display alongside a new orthographic instruction (e). The speaker gives the confederate verbal instructions (f) that the confederate follows by dragging the cube onto the next referent (g).

4 (meta-analysis) took longer than initially anticipated, and the total duration of the project was approximately 12 months.

The project OSF repository contains all the materials mentioned in this article and can be accessed at https://osf.io/3bmcp. The following sections report the criteria for sample size, data exclusions, data manipulations, and all the measures in the study.

Phase 1: recruitment of analysts and initial survey

An online landing page provided a general description of the project, including a short prerecorded slide show that summarizes the data set and research question (https://many-speech-analyses.github.io). The project was advertised via social media using mailing lists for linguistic and psychological societies and via word of mouth. Social-media advertising was accompanied by a short recruitment form. The target population comprised active speech-science researchers with a graduate/doctoral degree (or currently studying for a graduate/doctoral degree) in relevant disciplines. All individuals interested in participating were asked to complete a questionnaire detailing their familiarity with numerous analytic approaches common in the speech sciences. Researchers could choose to work independently or in small teams. For the sake of simplicity, we refer to a single researcher and teams both as **analysis teams**.[3] Recruitment for this project commenced after having received in-principle acceptance.

As outlined above, our primary aim is to assess the variability of the reported effects rather than the meta-analytic estimate of the investigated effect per se. To estimate the degree of uncertainty around effect variability as driven by number of teams, we ran a series of sample-size simulations with values of variability extracted from Silberzahn et al. (2018). The code is available at https://many-speech-analyses.github.io/many_analyses/scripts/r/simulations/simulations.[4] Variability among teams was operationalized as the standard deviation of the teams' reported effects from Silberzahn et al. (2018). We *z*-scored this variability prior to simulations to make it comparable to our study. For the mean of the teams' true standard deviation (*z* score of 0.68), the simulation indicates that the degree of uncertainty around the estimated teams' standard deviation will be below 1 *SD* at any sample size greater than 10 teams. Thus, to achieve our main goal (i.e., estimating variability among teams), we considered a minimum sample size of 10 teams as sufficient. Given the exploratory nature of our study, however, we sampled as many analysts as possible. We received initial expressions of interest to participate from more than 200 analysts, although there was a substantial dropout rate (see Results section).

After analysts submitted their analyses, we asked them to also function as peer reviewers. Each team had to review four other analyses. All analysts involved are coauthors on this article and participated in the collaborative process of producing the final manuscript. Informed consent was obtained as part of the intake form.

Phase 2: primary data analyses

The analysis teams registered for participation, and each of the analysts individually answered a demographic and expertise questionnaire. A PDF version of this and all other questionnaires are available at https://osf.io/h6z8w. The questionnaire collected information on the analysts' current position and self-estimated breadth and level of statistical expertise and acoustic-analysis skills. We then requested that they answer the following research question: "Do speakers acoustically modify utterances to signal atypical word combinations?" To do so, they were given the data generated by the experiment described in The Data Set section above. Data included the audio recordings with corresponding time-aligned transcriptions in the form of Praat TextGrid files. These files can be found at https://osf.io/5agn9.

Once their analysis was complete, they answered a structured questionnaire that provided information about their analysis technique, an explanation of their analytic choices, their quantitative results, and a statement describing their conclusions. They also uploaded their analysis files (including the additionally derived data and text files that were used to extract and preprocess the acoustic data), their analysis code (if applicable), and a detailed journal-ready analysis section.

Phase 3: peer review of analyses

The analyses from each team were evaluated by four different teams who functioned as peer reviewers. Each peer reviewer was randomly assigned to analyses from at least four analysis teams. Reviewers evaluated the methods of each of their assigned analyses one at a time in a sequence determined by the initiating authors. The sequences were systematically assigned so that, if possible, each analysis was allocated to each position in the sequence for at least one reviewer.

The process for a single reviewer was as follows. First, the reviewer received a description of the methods of a single analysis. This included the narrative methods and results sections, the analysis team's answers to the questionnaire regarding their methods, including the analysis code and data set. The reviewer was then asked in an online questionnaire to rate both the acoustic and statistical analyses and to provide an overall rating using a

scale of 0 to100. To help reviewers calibrate their rating, they were given the following guidelines:

- 100 = perfect analysis with no conceivable improvements from the reviewer
- 75 = imperfect analysis but the needed changes are unlikely to dramatically alter the final interpretation
- 50 = flawed analysis likely to produce either an unreliable estimate of the relationship or an overprecise estimate of uncertainty
- 25 = flawed analysis likely to produce an unreliable estimate of the relationship and an overprecise estimate of uncertainty
- 0 = dangerously misleading analysis certain to produce both an estimate that is wrong and a substantially overprecise estimate of uncertainty that places undue confidence in the incorrect estimate

The reviewers were also given the option to include further comments in a text box for each of the three ratings.

After submitting the review, a methods section from a second analysis was made available to the reviewer. This same sequence was followed until all analyses allocated to a given reviewer were provided and reviewed.[5]

Phase 4: evaluating variation

The initiating authors (S. Coretta, J. V. Casillas, and T. B. Roettger) conducted the analyses outlined in this section. We did not conduct confirmatory tests of any a priori hypotheses. We consider our analyses exploratory.

Descriptive statistics. We calculated summary statistics describing variation among analyses, including (a) the nature and number of acoustic measures (e.g., f_0 or duration), (b) the operationalization and the temporal domain of measurement (e.g., mean of an interval or value at a specified point in time), (c) the nature and number of model parameters for both fixed and random effects (if applicable), (d) the nature and reasoning behind inferential assessments (e.g., dichotomous decision based on p values, ordinal decision based on a Bayes factor), as well as the (e) mean, (f) standard deviation, and (g) range of the standardized effect sizes (see the next section for the standardization procedure). These summary statistics are reported in the Results section.

Meta-analytic estimation. We investigated the variability in **reported effect sizes** using Bayesian meta-analytic techniques. As the measure of variability, we took the **meta-analytic group-level standard deviation** (σ_{α_t}; see below), where each analysis team represents a group. As we detail in the Results section, we have also run further nonpreregistered analyses. For these we refer the reader to that section, while we describe only the preregistered analyses in the following paragraphs.

On the basis of the common practices currently in place within the field, we anticipated that researchers would use multilevel regression models; thus, common measurements of effect size, such as Cohen's d, might have been inappropriate. Furthermore, Aczel et al. (2021) suggested that directly asking analysts to report standardized effect sizes could bias the choice of analyses toward types that more straightforwardly return a standardized effect. Because the variables used by the analysis teams might have substantially differed in their measurement scales (e.g., Hertz for frequency vs. milliseconds for duration), which was indeed the case, we standardized all reported effects by refitting each reported model with centered and scaled continuous variables (z scores, i.e., the observed values subtracted from the mean divided by the standard deviation) and sum-coded factor variables. Each **standardized model** was fitted as a Bayesian regression model with Stan (Version 2.26.0; Stan Development Team, 2021), RStan (Version 2.21.2; Stan Development Team, 2020), and brms (Bürkner, 2017) in R (Version 4.2.1; R Core Team, 2020). Model refitting also constituted a way of validating the reported analyses, a step recommended by Aczel et al. (2021). Details about the refitting procedure can be found at https://many-speech-analyses.github.io/many_analyses/scripts/r/04_refit_workflow. Relative to the registered protocol, we made minor changes to the refitting procedure, specifically file and variable naming conventions and the use of treatment contrasts instead of sum coding. All models converged ($\hat{R}$ was approximately 1). Of the models with divergent transitions (n = 10), the number of divergences ranged from 1 to 156 (156 represents 3.9% of the total number of samples), which the authors deemed not to be problematic.

The coefficients of the critical predictors (i.e., critical according to the analysis teams' self-reported inferential criteria) obtained from the standardized models were used as the **standardized effect size** (η_i) of each reported model. Moreover, to account for the differing degree of uncertainty around each standardized effect size, we used the standard deviation of each standardized effect size as the **standardized standard error** (se_i). This enabled us to fit a so-called measurement-error model, in which both the standardized effect sizes and their respective standard errors are entered in the **meta-analytic model**. As a desired consequence, effect sizes with a greater standard error are weighted less than those with a smaller standard error in the meta-analytic calculations.

After having obtained the standardized effect sizes η_i with related standard errors se_i, for each critical predictor

in each reported model, we conducted a **Bayesian random-effects meta-analysis** using a multilevel (intercept-only) regression model. The outcome variable was the set of standardized effect sizes η_i. The likelihood of η_i was assumed to correspond to a normal distribution (Knight, 2000). The analysis teams were entered as a group-level effect (i.e., `(1 | team)`, called "random effect" in the frequentist literature). The standard errors se_i were included as the standard deviation of η_i to fit a measurement-error model, as discussed above. We used regularizing weakly informative priors for the intercept α (*Normal*(0,1)) and for the group-level standard deviation σ_{α_i} (*HalfCauchy*(0,1)). We fit this model with four chains of Hamiltonian Monte Carlo sampling for the estimation of the joint posterior distribution using the No U-Turn Sampler as implemented in Stan (Team, 2021), and 4,000 iterations (2,000 for warm-up) per chain, distributed across eight processing cores and two threads in within-chain parallelization. The model did not incur any divergent transitions ($\hat{R}$ was not greater than 1), and the estimated sample sizes were sufficient. The code used to run the model can be found at https://many-speech-analyses.github.io/many_analyses/scripts/r/06_meta-analysis_prereg.

The posterior distribution of the population-level intercept α allowed us to estimate the range of probable values of the standardized effect size $\hat{\eta}$. The posterior distribution further allowed us to investigate the effect of a set of analytic and researcher-related predictors, as detailed in the next section. Crucially, the posterior distribution of the group-level standard deviation σ_{α_i} (i.e., the standard deviation of the group-level effect of team) allowed us to quantify the degree of variation between the teams' analyses on a standardized scale.

Analytic and researcher-related predictors affecting effect sizes. As a second step, we investigated the extent to which the individual standardized effect sizes are affected by a series of analytic and researcher-related predictors.

Analytic predictors. We estimated the influence of the following predictors related to the analytic characteristics of each team's reported analysis:

- *Measure of uniqueness* of individual analyses for the set of predictors in each model (numeric)
- *Number of models* the teams reported to have run [numeric]
- *Major dimension* that has been measured to answer the research question [categorical]
- *Temporal window* that the measurement is taken over [categorical]
- *Average peer-review rating*, as the mean of the overall peer-review ratings for each analysis [numeric]

Following Parker et al. (2020), the measure of uniqueness of predictors was assessed by the Sørensen-Dice Index (SDI; Dice, 1945; Sørensen, 1948). The SDI is an index typically used in ecology research to compare species composition across sites. It is a distance measure similar to Euclidean distance measures but is more sensitive to more heterogeneous data sets and deemphasizes outliers. For our purposes, we treated predictors as *species* and individual analyses as *sites*. For each pair of analyses (X,Y; across and within teams), the SDI was obtained using the following formula:

$$\mathrm{SDI} = \frac{2\,|\,X \bigcap Y\,|}{|\,X\,| + |\,Y\,|}$$

where $|\,X \bigcap Y\,|$ is the number of variables common to both models in the pair, and $|\,X\,| + |\,Y\,|$ is the sum of the number of variables that occur in each model. For example, if two pairs of models differ in either only one predictor (e.g., DV ~ typicality vs. DV ~ typicality + trial) or in two predictors (e.g., DV ~ typicality vs. DV ~ typicality + trial + speech rate), the latter model pair would exhibit a larger SDI than the former. To generate a unique SDI for each analysis team, we calculated the average of all pairwise SDIs for all pairs of analyses using the `beta.pair()` function in the betapart R package (Baselga et al., 2020).

The major measurement dimension of each analysis was categorized according to the following possible groups: duration, intensity, f_0, other spectral properties (e.g., frequency, center of gravity, harmonics difference), and other measures (e.g., derived measures such as principal components, vowel dispersion). The temporal window that the measurement is taken over is defined by the target linguistic unit. We assume the following relevant linguistic units: *segment*, *syllable*, *word*, *phrase*, *sentence*. Because each analysis received more than one peer-review rating, we calculated the mean rating and its standard deviation for each. These were entered in the model formula as a measurement-error term (`me(mean, sd)` in brms).

Researcher-related factors. We also included the following predictors:

- *Research experience* as the elapsed time from receiving the PhD; negative values indicate that the person is a student or graduate student [numeric]
- *Initial belief* in the presence of an effect of atypical noun-adjective pairs on acoustics, as answered during the intake questionnaire [numeric]

To obtain an aggregated research-experience score and initial-belief score for each team on the basis of

members' individual scores, we calculated the mean and standard deviation of these predictors for each team. These were entered in the model formula as a measurement-error term (`me(mean, sd)` in brms). The expedient of using a measurement-error term (which includes the teams' standard deviation) ensures information about within-team variance is not lost (which would be the case if including the mean only).

We had initially planned to also include a measure of conservativeness of the model specification, as the number of random/group-level effects included and the number of post hoc changes to the acoustic measurements the teams reported to have carried out. When fitting the model, we realized that the measure of conservativeness is related to the standard error of the estimates (i.e., more group-level effects = higher standard error). Moreover, there was no team that declared to have made post hoc changes to the analyses; thus, we decided against including these two preregistered predictors in the model.

Model specification. The model was fitted as a measurement-error model, with the predictors detailed in the preceding paragraphs. The outcome variables of the model were the standardized effect sizes and related standard deviation.

A normal distribution was used as the likelihood function of $\alpha_{t[i]}$. The mean of $\alpha_{t[i]}$ was modeled on the basis of the overall intercept β and on the coefficients of each predictor. The numeric predictors were centered and scaled and the categorical predictors were sum-coded. We used a normal distribution with $M = 0$ and $SD = 1$ as the prior for the intercept and the predictors. The model was run with the same settings as with the meta-analytic model. The code used to run the model can be found at https://many-speech-analyses.github.io/many_analyses/scripts/r/06_meta-analysis_prereg.

Data management. All relevant data, code, and materials have been publicly archived on OSF (https://osf.io/3bmcp). Archived data include the original data set distributed to all analysts, any edited versions of the data analyzed by individual teams, and the data we analyzed with our meta-analyses, which include the standardized effect sizes, the statistics describing variation in model structure among analysis teams, and the anonymized answers to our questionnaires of analysts. Similarly, we archived both the analysis code used for each individual analysis and the code from our meta-analyses. We also archived copies of our survey instruments from analysts and peer reviewers. Further documents concerning the collaborative editing of the registered report can be found at https://drive.google.com/drive/folders/1-DOcj1qtEkvWfzu_FrsxkIGfPS0DyLXB?usp=sharing.

We excluded from our synthesis any individual analysis submitted after peer review (Phase 3) or those unaccompanied by analysis files, without which it was not possible to follow the research protocol. We also excluded any individual analysis that did not produce an outcome that could be interpreted as an answer to our primary question. We also did not include analyses for which we could not extract standardized effect sizes. For a list of exclusion criteria, see the Descriptive Statistics section below.

Phase 5: collaborative write-up of manuscript

The initiating authors discussed the limitations, results, and implications of the study and collaborated with the analysts on writing the final manuscript for review as a Stage 2 registered report.[6]

Results

This section is divided into three parts. We first provide a statistical description of team composition, nature of acoustic analyses and statistical approaches, and peer-review ratings. Second, we report the results of the meta-analytic model, focusing on between-team and between-model variability. Finally, we present the analysis of the effect of analytic and researcher-related predictors on the meta-analytic effect. The research compendium of the study, containing all the code and data presented here, can be found at https://osf.io/3bmcp. An interactive web application that allows the interested reader to explore the data set is available at https://many-speech-analyses.github.io/shiny.

Descriptive statistics

In the following sections, we describe the characteristics of the analysis teams that participated in the study and the analytic approaches they adopted. An important aspect that emerges from the descriptive analysis is the large variation in analytic strategies.

Characteristics of analysis teams. Eighty-four teams initially signed up to participate in the study, comprising 211 analysts. Thirty-eight of the signed-up teams dropped out during the analysis phase.

Forty-six teams submitted their analyses by the established deadline. Only analyses from which it was possible to extract an effect size were included in the meta-analysis. Of the analyses submitted by the 46 teams, the initiating authors identified 33 teams with submissions meeting the criteria for inclusion in the meta-analytic model. Reasons for exclusion were use of

generalized additive models (four teams), which do not lend themselves easily to the meta-analytic methods used in this study; use of machine-learning techniques (three teams); use of typicality as the outcome variable/response (three teams); or use of other methods that returned statistics that could not be included in the meta-analytic model. Note that due to the unforeseen variability across teams, the latter exclusion criteria were not preregistered and were applied after having seen all analytic strategies.

In what follows, we describe the characteristics of those teams whose analyses were included in the meta-analytic model. A complete summary of all the analyses from the 46 submitting teams is available at https://many-speech-analyses.github.io/many_analyses/RR_manuscript/supplementary_materials.pdf.

The included analyses were provided by 33 teams, comprising 120 analysts, with a median of 3.0 individuals per team. Upon sign-up, we collected background information from each analyst through the intake form, which was administered during Phase 1 before the data were released to the teams. Analysts had a median of 5.4 years of experience after completing their PhD, ranging from −3.8 years, that is, PhD students (or less experienced) to 12.4 years, suggesting that, on average, analysts were experienced researchers. The analysts' prior belief in the effect under investigation, on a scale from 0 to 100, ranged from 46.4 to 92.0 with a median of 70.0. We take this to suggest that, overall, analysts had a rather high positive prior belief in the investigated relationship between acoustics and word-combination typicality.

At the end of Phase 2 (primary data analysis), the teams had submitted a total of 115 individual models (including 192 critical model coefficients, given that some models returned more than one critical coefficient) to answer the research question, with a median of three models per team. Table 1 provides a summary of the contributing teams and their analyses.

Table 1. Descriptive Statistics of Teams, Acoustic Analyses, and Statistical Analyses Included in the Meta-Analysis

Team characteristics, range (*Mdn*)	
Team size	1.0–12.0 (3.0)
Years after PhD	−3.8–12.4 (5.4)
Prior belief	46.4–92.0 (70.0)
Acoustic analysis peer rating	41.2–88.3 (73.8)
Statistical analysis peer rating	33.0–93.3 (73.2)
Overall peer rating	39.0–88.7 (70.8)
Acoustic analyses, *n* (%)	
Outcome	
f_0	44 (37)
Duration	39 (33)
Intensity	18 (15)
Formants	15 (13)
Other	3 (3)
Temporal window	
Segment	54 (46)
Word	53 (45)
Sentence	4 (3)
Phrase	3 (3)
Other	4 (3)
Typicality operationalization	
Categorical	82 (69)
Continuous (mean)	33 (28)
Continuous (median)	3 (3)
Statistical analyses	
Framework	
Frequentist	100 (84)
Bayesian	19 (16)
Model	
Linear	117 (98)
GAM	1 (1)
Other	1 (1)
N, range (Mdn)	
Models	1–16 (3)
Predictors	1–5 (2)
Random terms	1–10 (2)
Intercept	1–10 (2)
Slope	0–4 (0)

Note: The data set included analyses from 33 teams and 120 analysts.

Acoustic analysis. The analytic teams differed in their approach to the acoustic analysis of the speech signal, including choices related to specific acoustic measures, the temporal window used, and how the measures were transformed. Thirty-seven percent of the models used f0 as the outcome variable, 33% used a measure of duration, 13% used vowel formants, 15% intensity, and 3% other measures.

Forty-five percent of models used acoustic measures taken at the level of the segment (e.g., comparing the acoustic profile of a vowel), 45% from the word level (e.g., comparing the acoustic profile of *Banane*; "banana"), 3% at the level of the phrase (e.g., the noun phrase including determiner and adjective, e.g., "the green banana"), 3% from the whole sentence, and 3% used a different time window. On the basis of a coarse coding of how acoustic measures were operationalized, we found a total of 55 different measurement specifications. For example, if we considered those analyses that targeted f_0, we found that it was operationalized in many different ways, including the minimum, maximum, mean, and median, as a range in an interval or a ratio between two intervals. The measurement was sometimes taken from the interval of a vowel in the article, adjective, or noun; it was sometimes taken from the word interval of

the article, adjective, or noun; or it was taken from either the noun-phrase interval or the entire sentence. Some of these measures were normalized relative to other elements in the sentence or relative to the speaker.

Statistical analysis. The large decision space related to how the acoustic signal was measured is further expanded by the choices in the statistical analysis, including the chosen inferential framework, the type of model, and the model specification, including choice of predictors, interactions, and group-level effects.

The mean of the number of different predictors included in teams' models was 2 (defined as variables or columns in the data table). This means that, in addition to the critical predictor (typicality of the adjective-noun combinations), models had on average one additional predictor (range = 1–5). Possible information that was used as predictors included the information structure of the sentence, trial number, semantic dimensions of the referent, part of speech, and speaker gender.

The data given to the teams allowed them to operationalize the predictor of interest, word typicality, in different ways. Among the possible operationalizations, 69% of models contained typicality as a categorical variable (e.g., atypical vs. typical), 28% used a continuous typicality scale from 0 to 100 by calculating the mean typicality for each word combination as obtained from the norming study, whereas 3% of the models used the median typicality rating. Note that the design of the experiment alongside its description indicated that the experiment was designed to categorically operationalize typicality. This possibly explains the analysts' strong preference.

The majority of the models were run within a frequentist framework (84%). Sixteen percent were run within a Bayesian framework. Although teams almost exclusively used linear models to analyze their data (98%), teams differed drastically in how they accounted for dependencies within the data.

The data contain several dependencies between data points, with multiple data points coming from the same subject and with multiple data points being associated with the same adjective or noun. An appropriate way to account for this nonindependence is by using models that include so-called random or group-level effects (e.g., Gelman & Hill, 2006; Schielzeth & Forstmeier, 2009), variably known as mixed-effect, hierarchical, multilevel, or nested models (among other names). Nine percent of the linear models specified no random effects at all (without pooling their data), effectively ignoring these nonindependences (Hurlbert, 1984). Sixty-two percent specified random intercepts only, and 29% specified both random intercepts and random slopes to account for the nonindependence. On average, teams that specified random effects included 2.5 random terms in their models. Based on statistical framework, type of model, distribution family, fixed terms, and not including random effects, there were a total of 52 different model specifications.

When considering both acoustic and statistical analyses, we have found a total of 119 different analytic pipelines. In other words, each individual analysis submitted was unique. A sankey diagram illustrating the relationship between choices related to outcome, temporal window, and operationalization can be found at https://many-speech-analyses.github.io/many_analyses/RR_manuscript/supplementary_materials.pdf.

Our quantitative assessment did not include other degrees of freedom, all of which are additional sources of variation: Teams differed (a) with regard to how the acoustic signal was segmented, ranging from fully automated forced alignment with minimal manual correction to complete manual alignment performed by the analysts; (b) in whether the statistical analysis was based on a subset of the data or the whole data set; and (c) whether and if so how measurements were excluded on the basis of both qualitative (i.e., whether specific speech-production instances were excluded or not) and quantitative grounds (i.e., whether data were trimmed or not).

The question arises whether these unique analysis pipelines led to different conclusions. Thirteen of the 33 teams (39.4%) reported to have found at least one statistically reliable effect (based on the inferential criteria they specified). Of the 192 critical model coefficients, 45 were claimed to show a statistically reliable effect (23.4%).

Review ratings. Teams reviewed each others' acoustic and statistical analyses. The mean rating of the acoustic analyses, on a scale from 0 to 100, was 71.5 (*SD* = 13.5). The mean rating of the statistical analysis was 69.4 (*SD* = 15.9). For reference, as mentioned in the Method section, a score of 75 was defined as "an imperfect analysis but the needed changes are unlikely to dramatically alter the final interpretation," indicating that on average reviewers judged the provided analyses to be appropriate, although "imperfect."

Meta-analytic estimation

This section deals with the meta-analytic analysis of the results submitted by the teams. As discussed above, the analyses of only 33 teams out of all the submitted analysis were included in the meta-analytic model discussed here. First, we report on the between-team variability estimate (i.e., the meta-analytic group-level standard deviation σ_{α_t}), which is the focus of this study, followed by the meta-analytic estimate, that is, the intercept of

the meta-analytic model (i.e., the estimated effect of typicality on the acoustic production of adjective-noun combinations).

Between-team variability. The primary aim of this analysis was to assess the degree of between-team variability. As a measure of between-team variability, we chose to use the meta-analytic group-level standard deviation (σ_{α_t}).

According to the preregistered meta-analytic model, the group-level standard deviation for teams was between 0.03 and 0.07 standard units at 95% credibility. In other words, the estimated range of variation across teams lies somewhere between ±0.06 (0.03 × 1.96) and ±0.13 (0.07 × 1.96) standard units with 95% credibility.

Non-preregistered. However, in our preregistration we did not take into account that teams might submit multiple analyses/models that, if unaccounted for, violate the independence assumption. Teams were explicitly instructed to submit only one effect size without enforcing it. As a result, some teams followed the instruction and submitted only one model, whereas others submitted multiple models. To account for this added layer of dependency, we ran a model with team and model ID nested within team as group-level effects (`(1|team) + (1|team:model_id)`), which allowed us to estimate both the between-team variation and the between-analysis variation. This analysis was not preregistered and should thus be interpreted with caution.[7]

The nested model yields a posterior 95% credible interval (CrI) for between-team variability of 0 to 0.04 standard units ($\beta = 0.02$, $SD = 0.01$), corresponding to a mean deviation range of about ±0 to ±0.1 standard units and 95% probability. The posterior 95% CrI for between-analysis variability (nested within teams) is 0.11 to 0.14 standard units ($\beta = 0.132$, $SD = 0.01$). For the sake of illustration, these would correspond to an estimate of between-model variability in segment and word durations that ranges between 7 to 14 ms for segments and between 7 and 33 ms for words at 95% credibility. We interpret these values in more detail in the Discussion section.

Taken together, the models suggest that the variability of reported effects between any model (within team or across) is substantially larger than the variability across individual teams. We return to this important observation later.

Meta-analytic intercept. After having assessed the variation between teams and analyses, we now turn to the meta-analytic estimate of the effect of typicality on the acoustic realization of sentences with adjective-noun combinations. The meta-analytic model estimates the range of probable values of the standardized effect size to be between −0.026 and 0.016 standard units (95% CrI, mean = −0.005). In other words, our best guess is that speakers might not encode typicality in the acoustic signal (e.g., by duration, f_0) or, if they do, they do so by a maximum of ±0.03 standard units.

Non-preregistered. As mentioned in the previous section, we ran an additional model using team and model ID nested within team as group-level effects. In this non-preregistered model, the meta-analytic intercept estimate was between −0.016 and 0.03 standard units (95% CrI, $\beta = 0.008$). This suggests that the acoustic measures of typical word combinations are 0.02 standard units lower to 0.03 standard units higher than the measures of atypical word combinations at 95% confidence. This result is qualitatively similar to the results obtained in the preregistered model.

The meta-analytic intercept conflates estimates from a variety of responses taken from very different places in the utterance (nouns, adjectives, determiners, entire phrases or sentences). This means that some of the effects on a particular response as observed in a specific location within the utterance might naturally be positive, whereas other might be negative, resulting in a meta-analytic intercept of about zero. We want to stress, however, that our focus is not on the meta-analytic intercept per se, but on the fact that a seemingly straightforward research question led to so many possible outcomes. We report more on this topic in the Discussion section.

Figure 3 illustrates the individual intercepts for critical typicality coefficients across models and teams, sorted in ascending order based on their mean. Given the nature and wide variety of acoustic operationalizations, there is no natural interpretation of the scale, so we cannot interpret the direction of estimates. When looking at the raw estimates and their variance (gray triangles and lines), it is striking how much estimates differed. Estimates ranged from −0.7 to 1.01 standard units.

Although the majority of model estimates and their uncertainty after shrinkage yields inconclusive results (i.e., are compatible with a point null hypothesis), there are 27 model estimates for which the 95% CrI does not contain zero (14%).

Analytic and researcher-related predictors

After having assessed the variability across teams and models, we now turn to estimating the impact of a series of predictors on the reported standardized effects. There is a large amount of variation between and within teams, raising the question as to whether we can explain some of this variation or whether it is purely idiosyncratic (Breznau et al., 2021).

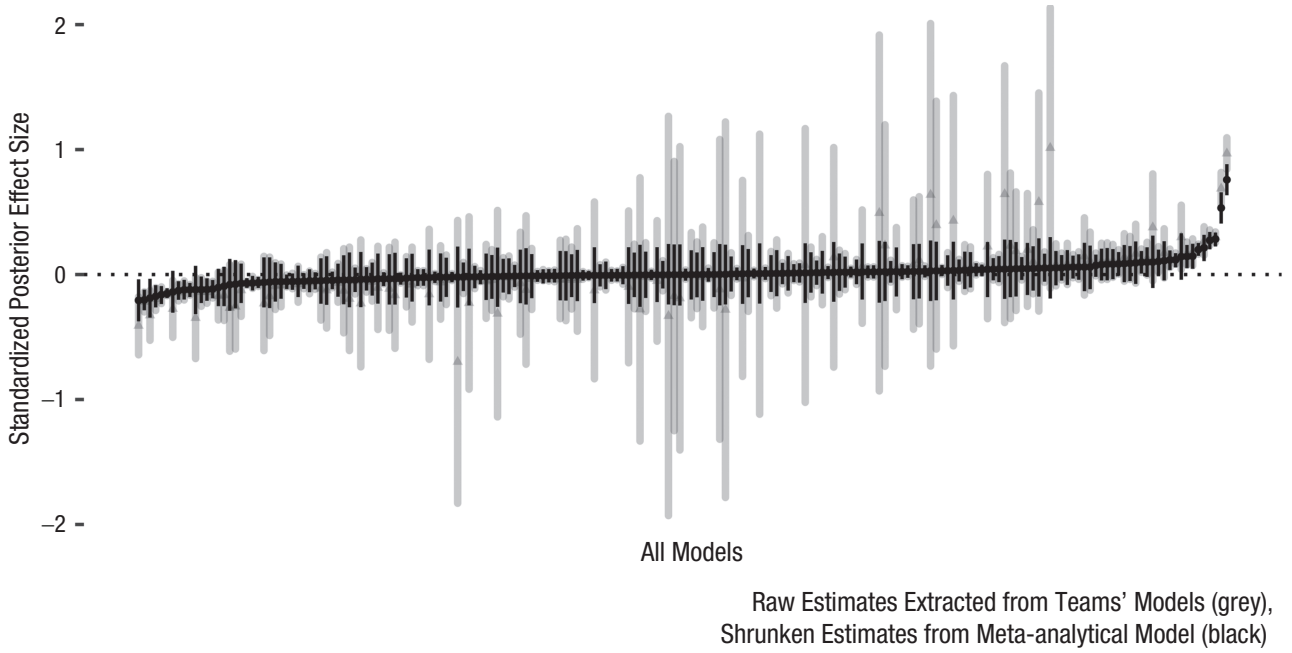

Fig. 3. Standardized effect sizes across all critical coefficients provided by the teams. Raw estimates are displayed in gray. Estimates after shrinkage as provided by the meta-analytic model are displayed in black.

We ran a model as described above. Figure 4c displays the coefficients for all predictors alongside their 80% and 95% CrIs. The model suggests that most team-specific predictors yielded very small deviations from the meta-analytic estimate, and their 95% CrIs included zero, leaving us highly uncertain about their direction. Neither analysts' prior beliefs in the phenomenon, $\beta = -0.01$, 95% CrI = [−0.04, 0.01], nor their seniority in terms of years after completing their PhD, $\beta = 0.01$, 95% CrI = [−0.02, 0.04], seem to have affected model estimates. Likewise, the evaluation of the quality of the analysis from their peers yielded a rather small effect magnitude, again characterized by large uncertainty, $\beta = 0.02$, 95% CrI = [−0.01, 0.05]. Interestingly, the model uniqueness, that is, how unique the choice and combination of predictors are, affected the analysts' estimate, with more unique models producing higher positive estimates, $\beta = 0.04$, 95% CrI = [0.02, 0.07].

Looking at the most important choices during measurement, both the acoustic parameter under investigation (e.g., f_0 or duration) and the choice of measurement window affected the results. Figures 4a and b display the posterior estimates for the measurement outcome (i.e., what acoustic dimension was measured; a) and measurement window (i.e., what is the unit over which the outcome was measured; b). If, on the one hand, an acoustic dimension related to f_0 was measured, estimates are lower than the meta-analytic estimate. If, on the other hand, duration was measured, estimates are higher than the meta-analytic estimate. Similarly, if acoustic parameters were measured across the entire sentence, estimates are lower than the meta-analytic estimate. In other words, depending on the choice of measurement and the measurement window, analysts might have arrived at different conclusions about how and if typicality is expressed acoustically.

It is due to the latter patterns that we need to interpret the results of the model with great caution. Because there are combinations of analytic choices that appear to systematically result in lower or higher estimates and the fact that predictors are not fully crossed (i.e., we do not have the same amount of data for all combinations of, e.g., outcome and measurement window), the estimates for certain predictors might be biased if predictors are collinear. This bias might be amplified by the fact that the scale has no natural way of being interpreted across all teams with different measurements cancelling each other out. We checked correlations between predictors, and although predictors do not seem to be highly collinear, the estimates might still be biased.

Discussion

Summary

We gave 46 analyst teams the same speech data set to answer the same research question: Do speakers acoustically modify utterances to signal atypical word combinations? To answer this question, teams had to interpret the research question by operationalizing constructs

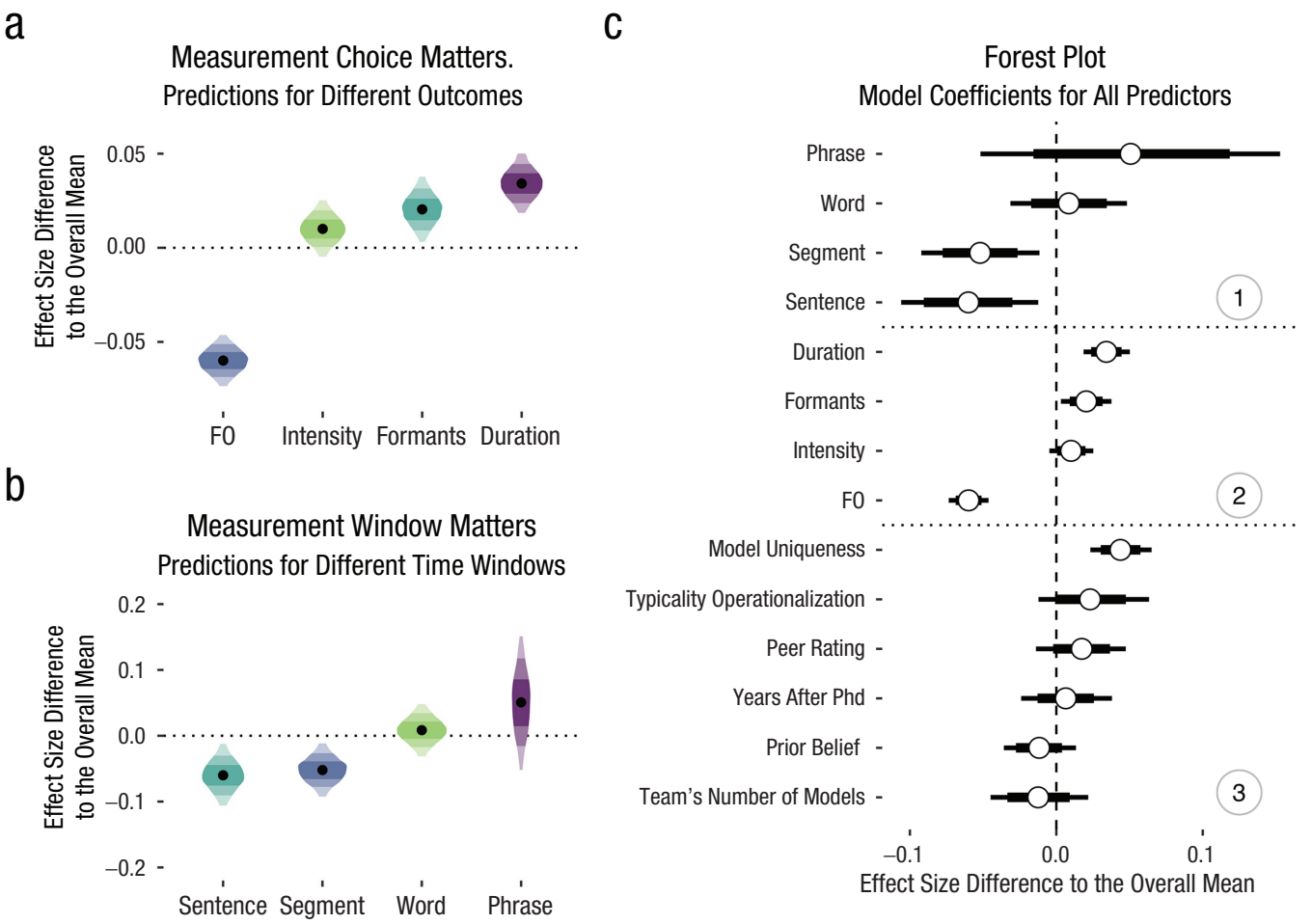

Fig. 4. The effects of analytic and researcher-related predictors on the reported standardized effect sizes. (a) Posterior samples for the four most frequent outcome variables; (b) posterior samples for the four most frequent temporal windows (black points = medians; shaded areas = 50/80/95% highest density intervals); and (c) mean posterior samples (white circles) and 80/95% credible intervals for all predictors grouped into predictors related to (1) temporal window, (2) outcome variable, and (3) team/analysis.

within multidimensional signals, operationalizing and choosing appropriate model predictors, and constructing appropriate statistical models. This complex process has led to a vast garden of forking paths, that is, to a wide range of combinations of possible analytic decisions. The submitted analyses exhibited at least 52 unique ways of operationalizing the acoustic signal alongside 55 unique ways of constructing the statistical model. By multiplying the numbers of acoustic and model specifications, there are in principle 2,860 possible unique combinations. Note that this is a conservative estimate of the number of possible analytic choices for our research question, ignoring many other degrees of freedom such as, for example, acoustic parameter extraction, outlier treatment, and transformations, all of which might have an impact on the final results (Breznau et al., 2021).

Different analysis paths led to different categorical conclusions with 39.4% of teams reported to have found at least one statistically reliable effect. To gain a better understanding of whether the observed quantitative variability can result in theoretically different claims, we will contextualize them in actual acoustic measures. We calculated the standard deviation of a selection of acoustic measurements, as submitted by the analysis teams: duration, f_0, and intensity, taken from different time windows. These standard deviations can be considered a coarse indication of the variability in the obtained acoustic measures. We can now use these values to interpret the meta-analytic estimates, which are in standardized units, by transforming the standardized units to measures of duration, f_0, and intensity (see Table 2 for examples of acoustic values grounding the estimated meta-analytic variation).[8]

For example, for those analyses that investigated the duration of vowels (e.g., the duration of the stressed vowel in *Banáne*), the reported duration measures exhibit standard deviations that range from 33.4 to 51.4 ms. These standard deviations allow us to convert the meta-analytic estimates into milliseconds by multiplying those values with the standard unit values of the

Table 2. Estimated 95% Credible Intervals of Deviation From the Meta-Analytic Effect in Acoustic Measures Based on the Lower and Upper Limits of the Between-Model Variation

Outcome	Temporal window	Lower	Upper	Unit
Duration	Segment	7–10.8	9.3–14.4	ms
Duration	Word	6.9–25.3	9.1–33.4	ms
f_0	Segment	0.9–9.4	1.2–12.4	Hz
f_0	Word	0.8–9.9	1.1–13.2	Hz
Intensity	Segment	0.7–1.5	0.9–2	dB
Intensity	Word	0.7–0.9	1–1.2	dB

meta-analytic estimates. The reported effect estimates from teams varied between −0.7 and 1.01 standard units, which corresponds to estimated segment-duration differences (for atypical vs. typical combinations) ranging from −23.34 to 33.84 ms. A more conservative approach is to convert the meta-analytic estimates of between-model variation, thus obtaining an estimate of between-model variability that ranges from 7.2 to 14.1 ms at 95% credibility. The calculation is thus: the minimum standard deviation of duration multiplied by the lower limit of the 95% CrI of the between-model variability estimate, times 1.96 to obtain a 95% CrI: 33.4 × 0.11 × 1.96 = 7.2 ms; the maximum standard deviation of duration multiplied by the upper limit of the 95% CrI of the between-model variability estimate, times 1.96: 51.4 × 0.14 × 1.96 = 14.1 ms.

Although this might not immediately strike one as highly variable, it crosses several theoretically relevant thresholds for perception and articulation: For example, the widely studied phenomenon of incomplete neutralization involves vowel-duration effects ranging from 7 to 15 ms (Nicenboim et al., 2018). This particular phenomenon has sparked long-lasting methodological and theoretical debates about the very nature of linguistic representations (Port & Leary, 2005) and has been replicated several times in both production and perception. Vowel duration differences within this range have also been reported across phenomena associated with segmental contrasts (Coretta, 2019), reduction phenomena (Nowak, 2006), and biomechanical reflexes of prominence (Mücke & Grice, 2014). Thus, variation between different analyst teams of 7.2 to 14.1 ms in one or the other direction can be theoretically relevant and might lead to opposing theoretical conclusions.

Although one might find it obvious that measuring different parts of the speech signal can lead to different results, the fact that analysts (and reviewers alike) considered all these data analytic pipelines valid ways of answering the same research question points to a lack of theoretical consensus on what parts of the speech signal correspond to what types of communicative functions. Importantly, even if analysts chose to measure more or less the same acoustic property within the same measurement window, they arrived at different estimates: For example, five teams measured f_0 in the noun and predicted f_0 on the basis of typicality as a categorical predictor. Their standardized effect estimates ranged from −0.35 to 0.19 standard deviations. Although these teams in principle measured the same thing, they differed in the analytical details of how f_0 was operationalized (i.e., mean, minimum, maximum, point or range) and how their statistical model was constructed (i.e., the number of predictors ranged from 1 to 2, and the number of random-effect terms ranged from 1 to 10). As shown by Breznau et al. (2021), even seemingly inconsequential analytical choices can affect conclusions in nontrivial ways.

The observed variation does not seem to be systematic. For example, variation between teams was not predicted by the analysts' prior expectations about the phenomenon. In fact, teams on average rated the plausibility of the effect as rather high before receiving access to the data. The observed variation was neither predicted by the analysts' experience in the field nor by the perceived quality of the analysis as judged by other teams. Analyses received overall high peer ratings for both the acoustic and the statistical analysis, suggesting that reviewers were generally satisfied with the other teams' approaches.

These findings are very much in line with previous crowdsourced projects that suggest variation between teams is neither driven by perceived quality of the analysis nor by analysts' biases or experience (e.g., Breznau et al., 2021; Silberzahn et al., 2018). Following Breznau et al. (2021), we are bound to conclude that "idiosyncratic uncertainty is a fundamental feature of the scientific process that is not easily explained by typically observed researcher characteristics or analytic decisions" (p. 9). Idiosyncratic variation across researchers might be a fact of life that we have to acknowledge and integrate into how we evaluate and present evidence.

Although properties of the teams did not seem to systematically affect the results, teams' estimates seem to highly depend on certain measurement choices. Human speech entails complex multidimensional

signals. Researchers need to make choices about what to measure, how to measure it, and which temporal unit to measure it in. Some of these choices seem to result in estimates in one direction, whereas others seem to result in estimates into another. For example, measurements related to f_0 tended to result in lower estimates, whereas measurements related to duration tended to yield higher estimates.

The asymmetry observed in the effect direction of different measurements can have several causes. First, there could be a true underlying relationship between typicality and the speech signal that manifests itself in some measures but not others and/or manifests itself negatively in one acoustic measure but positively in another.

Second and orthogonal to a possible true relationship, certain measurement choices might be associated with stronger expectations relative to the research question, which might lead to stronger researcher biases. Many analysts targeted measures related to f_0, likely because similar functional relationships such as information structure and predictability can be expressed by f_0 (e.g., Grice et al., 2017; Turnbull, 2017). Moreover, prior work has actually suggested a relationship between typicality and f_0 (e.g., Dimitrova et al., 2008, 2009). Participating analysts could have been aware of those findings, which might have, subconsciously or otherwise, nudged their choices in one particular direction.

Regardless of the cause of these systematic effects, we have to conclude that depending on the choice of how the speech signal is operationalized, researchers might find evidence for or against a theoretically relevant prediction. This conclusion is further supported by the fact that between-team variability was lower than between-model variability. This is an important observation when put into context of the fact that most teams submitted many different models. Teams submitted up to 16 different models to test for a possible relationship between typicality and the speech signal. The complexity of the speech signal lends itself to multiple approaches, but this plurality of hypothesis tests invites bias and can dramatically increase the rate of falsely claiming the presence of an effect (Roettger, 2019; Simmons et al., 2011). We of course are not arguing that exploratory analyses should not be used. Rather, we simply want to point out that if the theoretical underpinnings of the field were much clearer, different teams would have converged toward a limited set of analyses despite a less specific research question.

In relation to this aspect, one team coordinator decided to drop out of the project because of its approach being too top-down. The coordinator also expressed a preference to be able to explore and run a variety of descriptive analyses followed up with inferential statistics. We find that this attitude speaks to the main objective of the current study: investigate researchers' degrees of freedom in the speech sciences. Based on our personal experience with research in the field, it is common practice to test many different types of models, using many different types of measurements, to answer one research hypothesis. Although this is a valid way to explore data and generate new hypotheses, it is not suitable for hypothesis testing. When operating within the frequentist inferential framework, testing the same hypothesis with different dependent variables is known to increase the false-positive (Type-I error) rate. The well-established solution to this problem is to apply a correction for family-wise error (i.e., alpha correction). However, less clear-cut degrees of freedom, such as those observed in the current study, can not be corrected in a straightforward way. If left uncorrected, these degrees of freedom can nevertheless drastically inflate the false-positive rate, even if different choices are highly correlated (Roettger, 2019). Another possible outcome of analytic flexibility as seen in this study is selective reporting of those tests that yield a desirable outcome (John et al., 2012; Kerr, 1998; Simmons et al., 2011), while null results remain unreported (Rosenthal, 1979; Sterling, 1959). Fields such as the speech sciences that make theoretical advances based on multidimensional data should be aware of this flexibility and calibrate their confidence in empirical claims accordingly.

Looking at our results, one might argue (and this interpretation has been articulated by several teams during the collaborative write-up) that our sample of speech scientists actually converged on a qualitative conclusion; that is, there is no evidence for a relationship. However, if there truly was no underlying relationship, our results would suggest a concerning false-positive rate with 39.4% of teams reported to have found at least one statistically reliable effect. This rate is substantially higher than the conventionally accepted 5% false-positive rate in, for example, null-hypothesis significance testing frameworks. If, on the other hand, there actually was an underlying relationship, our results would suggest a concerning false-negative rate of 61.6%, with the majority of teams not detecting the effect. If the latter was true, the fact that the majority of teams arrived at a null result might also simply be a consequence of the sample size in the data set being too small to reliably detect an effect (which is unknown to us). Thus, we do not think that our study provides convincing evidence that speech researchers converged on the same qualitative answer to a broad research question.

Lessons for the methodological-reform movement

The current results point to important barriers to the successful accumulation of knowledge. The replication crisis has brought attention to scientific practices that lead to unreliable and biased claims in the literature

(Fidler & Wilcox, 2018; Vazire, 2017). One of the suggested paths forward is for researchers to directly replicate previous studies more often (Camerer et al., 2018; Open Science Collaboration, 2015). Although we agree with the importance of direct replications, our study (and similar crowdsourced analyses before us) suggest that replicating more is simply not enough. There is only limited value in learning that a particular procedure is replicable if the idiosyncratic nature of the procedure itself might not yield a representative result relative to all possible procedures that could have been applied to the research question. Thus beyond a mere replication crisis, quantitative disciplines are going through an "inference crisis" (Rotello et al., 2015; Starns et al., 2019). As shown by the peer ratings of the analyses reported in this study, well-trained and experienced speech researchers not only applied completely different approaches to the same research question but also considered most of these alternative approaches acceptable. Being aware of this idiosyncratic variation between analysts should lead to more nuanced claims and a certain level of epistemic humility (for an overview of the concept, see Campbell, 1975).

A desired outcome of knowing that different but reasonable measurement choices or statistical approaches might lead to different interpretations of research data is to calibrate our (un)certainty in the strength of the collected evidence and, in turn, communicate that (un) certainty appropriately. The fact that the choice of measurement, measurement window, and predictor choice affect the answer to the research question further suggests that research assumptions and hypotheses should be formulated in much greater detail, particularly so in regard to how measurement systems (here, the acoustic signal) and underlying conceptual constructs (here, the phonetic expression of typicality) relate to each other.

We should ideally specify the link between conceptual construct and quantitative system—the "derivation chain" (Dubin, 1970; Meehl, 1990)—before data collection and analysis, including defining constructs and their relationship within the quantitative system, specifying auxiliary assumptions and boundary conditions, and defining target measurements, statistical expectations, and possible (and impossible) effect magnitudes. Without well-defined derivation chains, we "are not even wrong" (Scheel, 2022) because falsified expectations cannot tell us much about the conceptual constructs they are based on when the relationship between the two is underspecified. Some of the analysis teams explicitly recognized and acknowledged the need to formulate a more precise version of the research question by preregistering their planned data analysis pipeline. Preregistration, that is, a time-stamped document in which researchers specify how they plan to collect their data and/or how they plan to conduct their confirmatory analysis, can be a useful tool to safeguard researchers against the urge to explore many different analytical paths before choosing the one that, in hindsight, seems most justified. However, as long as the theoretical landscape does not allow for more precise hypotheses, the value of preregistration is limited and we need to find ways to appropriately calibrate the confidence in our claims.

Through sharing of materials, data, and statistical protocols, we can make our idiosyncratic choices transparent to others (Munafò et al. 2017; Vazire, 2017). Sharing further enables the evaluation and verification of underlying claims and allows for the evaluation of empirical, computational, and statistical reproducibility (LeBel et al., 2018). It allows for alternative analyses to establish analytic robustness (Steegen et al., 2016) and strengthens attempts to synthesize evidence via meta-analyses (e.g., Nicenboim et al., 2018). Given that minor procedural changes can sometimes drastically affect the final interpretation of the results (Breznau et al., 2021), we should ideally share a detailed documentation of the data-collection procedure, the measurement choices, the data extraction, and statistical analyses. Within fields that deal with speech data, open-source software that permits the extraction of acoustic parameters via reproducible scripts can help other researchers to trace back seemingly inconsequential choices during the measurement process (e.g., Praat: Boersma & Weenink, 2021; EMU: Winkelmann et al., 2017; the Montreal Forced Aligner: McAuliffe et al., 2017).

Making analytic pathways completely retraceable and preregistering them in advance does not change the fact that different analysts might apply different analytic approaches (preregistered or not). Crowdsourced projects such as the current one can shed light on the range of degrees of freedom during analysis and could possibly help produce a consensual estimated effect if the research hypothesis is specific enough. Crowdsourcing analyses is obviously not always feasible in terms of required resources and time but could be a consideration for claims that have large epistemological or practical consequences.

If we develop a good understanding of relevant analytic degrees of freedom, we could apply all conceivable analytic strategies and compare the results across all combinations of these choices. Such an analysis can provide insight into how much the conclusions change due to analytic choices as well as which choices have neglible or large impact on the result. This approach is called a "multiverse analysis" (e.g., Harder, 2020; Steegen et al., 2016) and has recently gained popularity across disciplines.

Finally, neither crowdsourcing nor multiverse analyses will guarantee that all relevant pathways are explored. Crowdsourcing is limited by the sampled analysts and their biases. Multiverse analyses are limited even further

by the group of researchers who define possible analytic pathways. Eventually, a mature scientific discipline needs to develop a set of detailed quantitative hypotheses of how conceptual constructs manifest themselves in the measured system, that is, in the present case how communicative pressures of certain functions are expressed in the acoustic signal. Possible tools to strengthen theoretical development relate to mathematically formalizing verbal expectations or using computational models (e.g., Devezer et al., 2021; Guest & Martin, 2021; Scheel et al., 2021; van Rooij & Blokpoel, 2020). Although conceptually promising, in their current state, such formalized models typically work in spaces that are much lower in dimensionality than the complex systems in which we measure. Thus, future research should spend resources on attempting to quantitatively relate the abstract theoretical space to the complex measurement space.

Caveats

Our study has several limitations that need to be considered when interpreting our results.

First, although the total number of analyses is larger than most earlier crowdsourcing projects, it is likely to be too small to reliably estimate the impact of certain predictors. Because predictors' values were not systematically distributed across teams, our estimates are characterized by large uncertainty.

Second, uncertainty is further inflated by the fact that the research question presented to the teams was vague, despite being of a kind normally found in the speech-science literature: Do speakers acoustically modify utterances to signal atypical word combinations? Interpreting the research question/hypothesis differently in terms of its statistical consequences has recently been shown to explain some variation between analysis teams in many-analyses projects (Auspurg & Brüderl, 2021). The analysts might also have tried to answer different specific manifestations of the research question that was given to them, leading to different choices down the line (e.g., whether speakers modify f_0 in atypical adjectives). It could be argued that some teams would have not specified such a vague research question to begin with, which would have reduced the possible degrees of freedom substantially. However, this very underspecification of research hypotheses in the field of speech science (and beyond; see Scheel, 2022) is very common. For example, researchers seem to have not yet agreed on how to acoustically measure cross-linguistically common phenomena such as word stress (e.g., Gordon & Roettger, 2017). Research on acoustic markers of clinical conditions such as depression and schizophrenia are often difficult to compare because of the wide variety of different acoustic measures used (e.g., Cummins et al., 2015; Parola et al., 2022).

Third, the design of this crowdsourced study has artificially inflated the variability between teams by encouraging anticoordination strategies. Teams knew that there would be other analyst teams and therefore might have chosen a "less canonical" analysis. Because analysts were guaranteed to become coauthors of a (in principle) guaranteed publication, such an anticoordination approach was not explicitly disincentivized.

Forth, our sample is an opportunity sample. We have advertised the project through online platforms that might have led to the exclusion of certain potential researcher groups. The sampling strategy also might have given access to researchers who were less experienced in particular aspects of the data analysis, possibly introducing uncommon analytic choices or poor-quality analyses. However, to our knowledge, neither the peer review among teams nor the information gathered through our questionnaires indicated any obvious cases of what one might consider incompetent analyses.

In light of both the observed large variability between teams, and possible sources of bias, a field can benefit from explicit positionality statements (e.g., Darwin Holmes, 2020; Fox et al., 2021; Jafar, 2018). Researchers do not analyze data in a vacuum. It is important to recognize and disclose one's positionality (i.e., a reflection about how educational background, social identity, power, experience, and context might influence researchers' approaches and interpretations). For example, the coordinating authors have engaged with meta-scientific research before and have been actively involved in methodological debates about scientific practices, including transparency and statistical methods. They have in the past used the lack of standardized analytic approaches as an argument for proposing behavior and policy changes in the field. This might have biased their own judgment during the analysis, which itself came with many researcher degrees of freedom. We hope we were able to make these degrees of freedom as well as the timing and reasoning of these analytic choices at least detectable, and we invite other researchers to reanalyze our data and try to replicate our results using a different research question.

Finally, the current study focused on a particular phenomenon within the speech sciences using a speech production data set with very specific properties. The generalizability of our findings to other disciplines, as well as to other subdisciplines of the language sciences specifically, is, of course, limited. We focused on quantitative analyses that require the operationalization of a multidimensional signal in an artificial elicitation situation (laboratory speech). Although we do believe that our qualitative conclusions hold across fields exhibiting similar methodologies, the detailed quantitative results will only be able to directly inform similar disciplines that work with speech or audio/video signals. This is an

important point to make because cognitive sciences in general, and the language sciences in particular, have many research areas that are based on qualitative methods (Haven & Van Grootel, 2019). It is conceivable that the discussed issues apply differently or not at all to qualitative data analyses.

Conclusion

Several recent studies have highlighted the large degree of analytic flexibility in data analysis. When many different analysts have to analyze the same data set to answer the same research question, analysts differ in how they approach this task, leading to both different qualitative answers (i.e., whether there is evidence for a relationship or not) and different effect magnitudes. This is concerning because it can lead to substantially different conclusions based on the same data set, a state of affairs that can generate biased inferential decisions and might weaken confidence in the published literature. More specifically, research commonly proceeds based on publications by one research team at a time. If we imagine a situation in which any of the 46 teams could have been *the* team publishing a study on this topic, it is immediately clear that that single study is just a very limited view. In light of this we want to stress that the field has to quickly move from one-off studies to collaborative approaches such as the one used here and to more frequent replication attempts, for example, by incentivizing replication through dedicated funding and editorial policies, among others.

Going beyond previous empirical studies, the current article looked at many analyses of speech data. Speech is a multidimensional signal that allows for great flexibility because it lends itself to a variety of possible operationalizations. In this study, 46 teams of speech scientists analyzed the same data set. Analytic approaches differed vastly in terms of their operationalization of key constructs, as well as their statistical analyses. Given the observed variability, conservative estimates of the sheer number of possible analytic paths for this research question lies in the thousands. Quantitatively, the between-team and between-model variation of estimates crosses important theoretical thresholds as to what constitutes communicative, cognitive, or biomechanical values.

In line with previous findings, neither the perceived quality of analyses nor the experience or prior beliefs of teams explained the observed variation. Importantly, however, we found some evidence for systematic effects on teams' estimates based on what and how they measured the speech signal. This result, taken together with the meaningful between-model variation and the tendency to test the research question on multiple outcome variables, suggests that a vast plurality of acceptable approaches is expected to frequently lead to different conclusions. We suggest that fields that use multidimensional data need to acknowledge these degrees of freedom, consider crowdsourcing and multiverse analyses when evaluating epistemologically or practically important phenomena, and strengthen the link between theoretical predictions and the measurement system by means of mathematical formalization and computational modeling.

Appendix

Glossary

- **Analysis team:** team of analysts or single analyst
- **Reported effect sizes:** effect sizes reported by each analysis team
- **Standardized model:** Bayesian refit of the team's model
- **Standardized effect sizes:** (η_i) effect sizes returned by the standardized models
- **Standardized standard error:** (se_i) standard deviation of the standardized effect sizes
- **Bayesian random-effects meta-analysis and meta-analytic model:** multilevel intercept-only regression model for meta-analysis
- **Meta-analytic group-level standard deviation:** (σ_{α_t}) standard deviation of the group-level effect of team returned by the meta-analytic model
- **Analytic and researcher-related predictors:** predictors used in the model that assess the effect of analytic and researcher-related factors on the standardized effects

Transparency

Action Editor: Julia Strand
Editor: David A. Sbarra
Author Contribution(s)
Stefano Coretta, Joseph Casillas, and Timo Roettger conceptualized and managed the project, ran the meta-analyses, and drafted both the Registered Report and the final manuscript. Simon Roessig, Michael Franke, and Timo Roettger collected the original data set. All other authors functioned as analysts of the data set and edited the final manuscript. Detailed author contributions are documented using the Contributor Roles Taxonomy (CRediT) and can be found here: https://github.com/many-speech-analyses/many_analyses/blob/main/figs/credit-taxonomy-all.png.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
S. Coretta was partially supported by European Research Council Grant No. 742289 (2017–2022; to J. Harrington); J. Al-Tamimi was partially supported by French Investissements d'Avenir - Labex EFL Grant ANR-10-LABX-0083,

contributing to IdEx Université Paris Cité Grant ANR-18-IDEX-0001; G. Moroz and S. Carraturo received funding from the Basic Research Program at the National Research University Higher School of Economics; I. Urrestarazu-Porta received funding from French National Research Agency Grant ANR-20-CE27-0007 and Spanish Ministry of Science and Innovation Grant PID2020-118445GB-I00; C. Kaland, M. Grice, F. Cangemi, M. Lialiou, M. Spaniol, and S. Wehrle received funding from German Research Foundation Grant 281511265-SFB 1252; and N. R. Benway was supported in part through computational resources provided by Syracuse University Grants NSF ACI-1341006 and NSF ACI-1541396.

ORCID iDs

Stefano Coretta https://orcid.org/0000-0001-9627-5532
Ali H. Al-Hoorie https://orcid.org/0000-0003-3810-5978
Najd E. Alotaibi https://orcid.org/0000-0003-3306-5081
George Bailey https://orcid.org/0000-0001-5137-8394
Nina R. Benway https://orcid.org/0000-0003-0955-9495
Aaron Braver https://orcid.org/0000-0001-8532-0473
Erin M. Buchanan https://orcid.org/0000-0002-9689-4189
Andrés Buxó-Lugo https://orcid.org/0000-0001-8274-035X
Francesco Cangemi https://orcid.org/0000-0003-1016-5178
Carissa A. Diantoro https://orcid.org/0000-0001-5286-0860
Shiloh Drake https://orcid.org/0000-0003-2247-2052
Ander Egurtzegi https://orcid.org/0000-0002-3451-323X
Mahmoud M. Elsherif https://orcid.org/0000-0002-0540-3998
Sara Finley https://orcid.org/0000-0002-7090-8108
Robert Fromont https://orcid.org/0000-0001-5271-5487
Pia Greca https://orcid.org/0000-0003-3491-1446
Amelia J. Gully https://orcid.org/0000-0002-8600-121X
Matthew C. Kelley https://orcid.org/0000-0002-7218-5599
Ghada Khattab https://orcid.org/0000-0002-8451-8135
Tomas O. Lentz https://orcid.org/0000-0001-8307-9639
Maria Lialiou https://orcid.org/0000-0002-9788-1443
Justin J. H. Lo https://orcid.org/0000-0003-0115-6982
Julio Cesar Lopez Otero https://orcid.org/0000-0001-6678-722X
Mridhula Murali https://orcid.org/0000-0001-5450-6419
Ladislas Nalborczyk https://orcid.org/0000-0002-7419-9855
Heather M. Offerman https://orcid.org/0000-0001-7535-2279
Maud Pélissier https://orcid.org/0000-0001-6639-9665
Scott J. Perry https://orcid.org/0000-0003-0400-0625
Michael Proctor https://orcid.org/0000-0002-3083-6859
Erik Schleef https://orcid.org/0000-0001-6636-1085
Joseph A. Stanley https://orcid.org/0000-0002-9185-0048
Benjamin V. Tucker https://orcid.org/0000-0001-8965-7890
Kingsley O. Ugwuanyi https://orcid.org/0000-0002-6480-0352
Iñigo Urrestarazu-Porta https://orcid.org/0000-0002-4606-3977
Emiel van Miltenburg https://orcid.org/0000-0002-7143-8961
Simon Wehrle https://orcid.org/0000-0001-9715-9541
Anna Wood https://orcid.org/0000-0002-4423-8219
Chenzi Xu https://orcid.org/0000-0001-6506-4513
Cong Zhang https://orcid.org/0000-0002-2561-2113
Timo B. Roettger https://orcid.org/0000-0003-1400-2739

Acknowledgments

We would like to thank the editors David Sbarra and Julia Strand, Matthew Goldrick, and two anonymous reviewers for their insightful comments and suggestions. The initiating authors (S. Coretta, J. V. Casillas, and T. B. Roettger) wish to express immense gratitude to all the other authors, without which this project would not have come to be.

Notes

1. Results of this research project were neither published nor publicly presented and are stored on a private OSF repository.
2. A detailed description of the norming and production studies from the Prosodic Encoding of Redundant Referring Expressions project, which was given to the analysts with the data set, can be found in methods_norm_prod.pdf at https://bit.ly/3Ahawc7.
3. Terms in bold are included with their definition in the glossary at the end of the paper for the reader's convenience.
4. Cached model outputs can be found at https://osf.io/wds2m.
5. Initially we planned to present simultaneously all four (or more) methods sections to each reviewer after the fourth round, with the option to revise their original ratings and provide an explanation. Ultimately, we decided to skip this step because of time constraints.
6. The comment history can be found at https://docs.google.com/document/d/1CFgRo93mRgifpuFOuQE3vNBeMW-H7ps9eD–vxH-6CQ/edit?usp=sharing.
7. Note that before fitting this model, we fitted a separate one in which model ID was the only (nonnested) group-level effect. The estimated group-level effect of model ID is identical to that of the nested model, so we do not discuss it further.
8. Note that these categories necessarily refer to a variegated set of measures; for example, the domain "word" includes words that differed along several dimensions, including their length and metrical structure.

References

Aczel, B., Szaszi, B., Nilsonne, G., Van den Akker, O., Albers, C. J., van Assen, M. A. L. M., Bastiaansen, J. A., Benjamin, D. J., Boehm, U., Botvinik-Nezer, R., Bringmann, L., Busch, N., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N., Donkin, C., van Doorn, J., . . . Wagenmakers, E.-J. (2021). *Guidance for multi-analyst studies*. OSF. https://doi.org/10.31222/osf.io/5ecnh

Arts, A., Maes, A., Noordman, L. G., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, *49*(3), 555–574.

Auspurg, K., & Brüderl, J. (2021). Has the credibility of the social sciences been credibly destroyed? reanalyzing the "many analysts, one data set" project. *Socius*, *7*, https://doi.org/10.1177/23780231211024421

Baselga, A., Orme, D., Villeger, S., De Bortoli, J., Leprieur, F., & Logez, M. (2020). *betapart: Partitioning beta diversity into turnover and nestedness components* (R package version 1.5.2). https://CRAN.R-project.org/package=betapart

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S. M., de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., . . . Bringmann, L. F. (2020). Time to get personal? the impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, Article 110211, https://doi.org/10.1016/j.jpsychores.2020.110211

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* [Computer software]. Praat. https://www.fon.hum.uva.nl/praat

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., . . . Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., Berthold, A., . . . Nguyen, H. H. V. (2021). *Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty*. OSF. https://doi.org/10.31222/osf.io/cd5j9

Brugger, P. (2001). From haunted brain to haunted science: A cognitive neuroscience view of paranormal and pseudoscientific thought. In J. Houran & R. Lange (Eds.), *Hauntings and poltergeists: Multidisciplinary perspectives* (pp. 195–213). McFarland & Company.

Burdin, R. S., Phillips-Bourass, S., Turnbull, R., Yasavul, M., Clopper, C. G., & Tonhauser, J. (2015). Variation in the prosody of focus in head-and head/edge-prominence languages. *Lingua*, *165*, 254–276. https://doi.org/10.1016/j.lingua.2014.10.001

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, D. T. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist*, *30*(12), 1103–1126.

Charles, S. J., Bartlett, J. E., Messick, K. J., Coleman, T. J., & Uzdavines, A. (2019). Researcher degrees of freedom in the psychology of religion. *The International Journal for the Psychology of Religion*, *29*(4), 230–245.

Coretta, S. (2019). An exploratory study of voicing-related differences in vowel duration as compensatory temporal adjustment in Italian and Polish. *Glossa: A Journal of General Linguistics*, *4*(1), 1–25.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49.

Darwin Holmes, A. G. (2020). Researcher positionality: A consideration of its influence and place in qualitative research—A new researcher guide. *Shanlax International Journal of Education*, *8*(4), 1–10. https://doi.org/10.34293/education.v8i4.3232

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591–621.

De Groot, A. D. (2014). *Thought and choice in chess* (Vol. 4). Walter de Gruyter GmbH & Co KG.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), Article 200805.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Dimitrova, D. V., Redeker, G., Egg, M., & Hoeks, J. C. (2008). Prosodic correlates of linguistic and extra-linguistic information in Dutch. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 2191–2196). Cognitive Science Society.

Dimitrova, D. V., Redeker, G., & Hoeks, J. C. (2009). Did you say a blue banana? The prosody of contrast and abnormality in Bulgarian and Dutch. In M. Uther, R. Moore, & S. Cox (Eds.), *Proceedings of Tenth Annual Conference of the International Speech Communication Association* (pp. 999–1002).

Dubin, R. (1970). Theory building. *Philosophy and Phenomenological Research*, *31*(2), 309–310.

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A. M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y. S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., . . . Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069.

Fidler, F., & Wilcox, J. (2018). *Reproducibility of scientific results*. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Winter 2018 ed.). Stanford University. https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under

uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288–299.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, *34*(4), 409–438.

Fox, J., Pearce, K. E., Massanari, A. L., Riles, J. M., Szulc Ranjit, Ł. Y. S., Trevisan, F., Soriano, C. R. R., Vitak, J., Arora, P., Ahn, S. J., Alper, M., Gambino, A., Gonzalez, C., Lynch, T., Williamson, L. D., & Gonzales, A. L. (2021). Open science, closed doors? Countering marginalization through an agenda for ethical, inclusive research in communication. *Journal of Communication*, *71*(5), 764–784.

Gatt, A., van Gompel, R. P., van Deemter, K., & Krahmer, E. (2011, July 20). *Are we Bayesian referring expression generators* [Paper presentation]. Proceedings of the Workshop on the Production of Referring Expressions (PRE-CogSci 2011), Boston, MA, United States.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis—A "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460–466.

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1), 1–11.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). Brill.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, *64*, 90–107.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Harder, J. A. (2020). The multiverse of methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, *15*(5), 1158–1177.

Haven, T. L., & Van Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229–244.

Hawkins, S., & Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics*, *32*(2), 199–231.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*(2), 187–211.

Jafar, A. J. N. (2018). What is positionality and should it be expressed in quantitative studies? *Emergency Medicine Journal*, *35*, 323–324. https://doi.org/10.1136/emermed-2017-207158

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), Article 20. https://doi.org/10.1525/collabra.158

Knight, K. (2000). *Mathematical statistics*. Chapman & Hall.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608–614.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E. J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., . . . Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479.

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

Lisker, L. (1977). Rapid versus rabid: A catalogue of acoustic features that may cue the distinction. *The Journal of the Acoustical Society of America*, *62*(S1), S77–S78.

Lisker, L. (1986). "voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*(1), 3–11.

Mati'c, D., & Wedgwood, D. (2013). The meanings of focus: The significance of an interpretation-based category in cross-linguistic analysis1. *Journal of Linguistics*, *49*(1), 127–163.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. In F. Lacerda (Ed.), *Proceedings of the Interspeech 2017* (pp. 498–502). https://doi.org/10.21437/Interspeech.2017-1386

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., et al. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31.

Mücke, D., & Grice, M. (2014). The effect of focus marking on supralaryngeal articulation–is it mediated by accentuation? *Journal of Phonetics*, *44*, 47–61.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. https://doi.org/10.1038/s41562-016-0021

Nicenboim, B., Roettger, T. B., & Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*, *70*, 39–55.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Niebuhr, O., d'Imperio, M., Fivela, B. G., & Cangemi, F. (2011). Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category. In W. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 120–123).

Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141.

Nowak, P. (2006). *Vowel reduction in Polish*. [PhD Dissertation, University of California, Berkeley]. https://escholarship.org/uc/item/1vh204j4

Ogden, R. (2004). Non-modal voice quality and turn-taking in Finnish. In E. Couper-Kuhlen & C. E. Ford (Eds.), *Sound patterns in interaction: Cross-linguistic studies from conversation* (pp. 29–62). John Benjamins.

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

Paraboni, I., Van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, *33*(2), 229–254.

Parker, T., Fraser, H., Nakagawa, S., Gould, E. B., Griffith, S., Vesk, P., & Fidler, F. (2020). Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology [Registed Report Stage 1 Protocol] (Version 1). *figshare*. https://doi.org/10.6084/m9.figshare.12034833.v1

Parola, A., Lin, J. M., Simonsen, A., Bliksted, V., Zhou, Y., Wang, H., Inoue, L., Koelkebeck, K., & Fusaroli, R. (2022). Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of NLP automated measures of coherence. *Schizophrenia Research*. Advance online publication. https://doi.org/10.1016/j.schres.2022.07.002

Port, R. F., & Leary, A. P. (2005). Against formal phonology. *Language*, *81*(4), 927–964.

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1). https://doi.org/10.5334/labphon.147

Roettger, T. B., Winter, B., & Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics*, *73*, 1–7.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, *1*(1), 75–116.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944–954.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*, Article 153. https://doi.org/10.3389/fpsyg.2016.00153

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), Article e2295. https://doi.org/10.1002/icd.2295

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., . . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, *5*(4), 1–34.

Stan Development Team. (2020). *RStan: The R interface to Stan* (Version 2.21.2). https://mc-stan.org

Stan Development Team. (2021). *Stan modeling language users guide and reference manual* (Version 2.26.0). https://mc-stan.org

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., . . . Wilson, J. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335–349. https://doi.org/10.1177/2515245919869583

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT Press. https://doi.org/10.1121/1.1327577

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Pearson.

Turnbull, R. (2017). The role of predictability in intonational variability. *Language and Speech*, *60*(1), 123–153.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van Heuven, V. J., Haan, J., Gussenhoven, C., & Warner, N. (2002). Temporal distribution of interrogativity markers in Dutch: A perceptual study. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology* (Vol. 7, pp. 61–86). Walter de Gruyter.

Van Heuven, V. J., & Van Zanten, E. (2005). Speech rate as a secondary prosodic characteristic of polarity questions in three languages. *Speech Communication*, *47*(1–2), 87–99.

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *Social Psychology*, *51*(5), 285–298.

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*(1), Article 1. https://doi.org/10.1525/collabra.74

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*, Article 935. https://doi.org/10.3389/fpsyg.2015.00935

White, L., Payne, E., & Mattys, S. L. (2009). Rhythmic and prosodic contrast in Venetan and Sicilian Italian. In M. Vigario, S. Frota, & M. J. Freitas (Eds.), *Phonetics and phonology: Interactions and interrelations* (pp. 137–158). John Benjamins.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, Article 1832. https://doi.org/10.3389/fpsyg.2016.01832

Winkelmann, R., Harrington, J., & Jänsch, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, *45*, 392–410.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967.