Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses

Stefano Coretta*[1, 2], Joseph V. Casillas[3], Timo B. Roettger[4]

Prepared using sagej.cls

Abstract

Recent empirical studies have highlighted the large degree of analytic flexibility in data analysis which 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 the present 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 find 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

[1]Department of Linguistics and English Language, University of Edinburgh, United Kingdom
[2]Institute of Phonetics and Speech Processing, Ludwig-Maximilian University Munich, Germany
[3]Department of Spanish and Portuguese, Rutgers University, New Brunswick, United States
[4]Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo, Norway

Corresponding author:
Timo B. Roettger
Email: timo.b.roettger@iln.uio.no

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

In order to effectively accumulate knowledge, science needs (i) to produce data that can be replicated using the original methods and (ii) to 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., Open Science Collaboration 2015; Camerer et al. 2018) 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: Simmons et al. 2011; Gelman and Loken 2014). 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; Tukey 1977; De Groot 2014). They are a curse because idiosyncratic choices can lead to categorically different interpretations, which 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., Silberzahn et al. 2018; Starns et al. 2019; Botvinik-Nezer et al. 2020). 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 like 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, in

other words, how to operationalize the phenomenon under investigation. This is not trivial, given the temporal extension of the acoustic signal and its complex structural composition.

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 goals are twofold: (i) our study conceptually replicates previous many-analyses projects, by probing the effects of different statistical analyses and by assessing the generalizability of published findings to other disciplines (here, the speech sciences); (ii) our study extends 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, for example 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. Humans are biased (e.g., Tversky and Kahneman 1974), e.g. they 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 (Sterling 1959; Koole and Lakens 2012), we often find ourselves exploring many possible analytic pipelines, but only reporting 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 (Simmons et al. 2011; Wagenmakers et al. 2012; Nosek and Lakens 2014) and in regard to its application in individual scientific fields (e.g. Wicherts et al. 2016; Charles et al. 2019; Roettger 2019). This notwithstanding, there are still many unknowns regarding the extent of analytic plurality in practice.

Consequently, a substantial body of published papers likely present overconfident interpretations of data and statistical results based on idiosyncratic analytic strategies (e.g., Simmons et al. 2011; Gelman and Loken 2014). 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. Miguel et al. 2014; Klein et al. 2018) and the advent of freely available and accessible infrastructures, such as those provided by the Open Science Framework (osf.io), critical re-analyses of published analytic strategies are still uncommon because data sharing remains rare (Wicherts et al. 2006).

Crowd-sourcing 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 twenty-nine 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, consequently, the results varied widely between teams.

Twenty teams (69%) found support for the hypothesis, and 9 (31%) did not. Out 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 byproduct of the scientific process and not necessarily driven by different levels of expertise or bias.

Several other recent studies 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 (see also Parker et al. 2020, and Botvinik-Nezer et al. (2020), regarding analytic flexibility in ecology and neuroimaging data, respectively). Bastiaansen et al. (2020) crowd-sourced 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 which might exhibit different degrees of theoretical underspecification. This is in line with Auspurg and Brüderl (2021) who re-analyzed the red card study mentioned above. The authors argue 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.

While these studies attested 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, where 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) discuss the cascading impact that different practices can have on psychometric research. The authors highlight, among others, 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 of 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 this data exponentiates the number of possible analytic approaches, thus further increasing analytic flexibility. In order to estimate this increased flexibility, the present 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 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 as 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 non-trivial 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, i.e. 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 which 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 and 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 sub-segmental intervals), spectral characteristics (intensity, formants, and spectral tilt), and measurements related to fundamental frequency (f0) (e.g., Gordon and Roettger 2017). Moving on to the expression of higher-level communicative functions, like 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 position, shape, and alignment of f0 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 vs. local pitch modulations (Van Heuven et al. 2002), as well as acoustic information within the temporal or spectral domain (e.g., Van Heuven and Van Zanten 2005). Extra-linguistic information, like 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 and 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* (see 2a and 2b).

(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 f0, 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

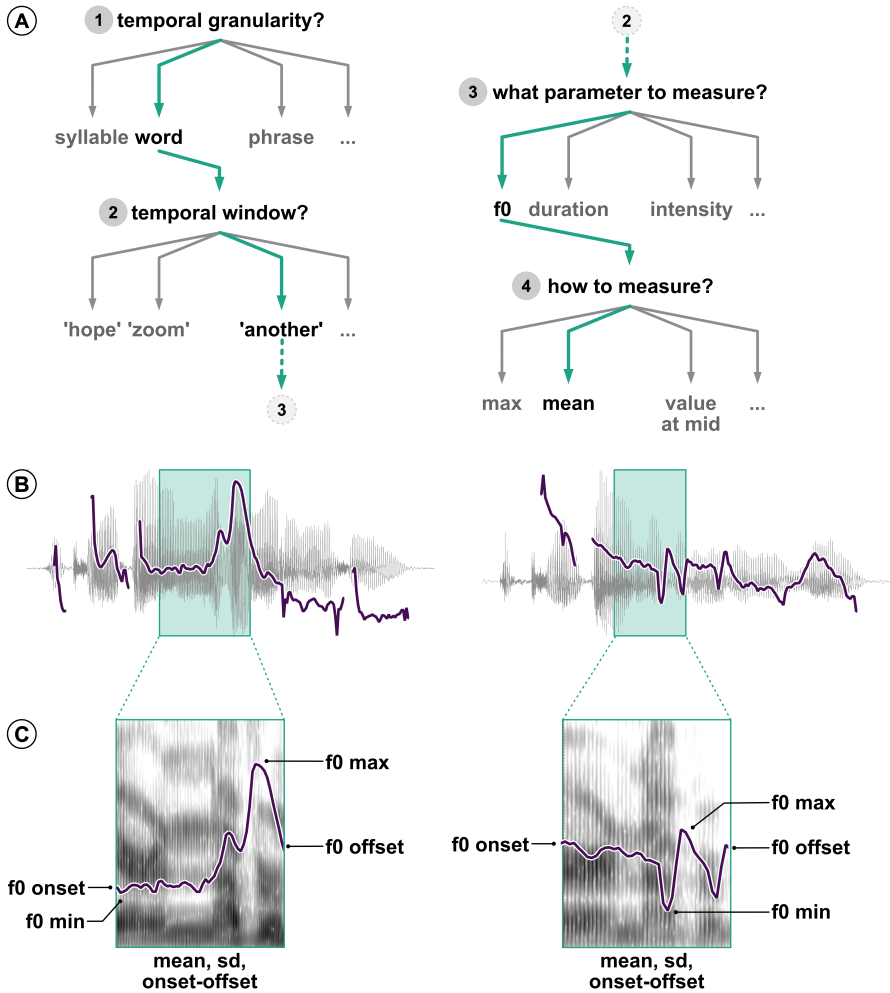

Figure 1. Illustration of the analytic flexibility associated with acoustic analyses. (A) An example of multiple possible and justifiable decisions when comparing two utterances; (B) Waveform and f0 track of the utterances *I can't bear ANOTHER meeting on Zoom* and *I can't bear another meeting on ZOOM*. The green boxes mark the word *another* in both sentences; (C) Spectrogram and f0 track of the word *another*, exemplifying possible operationalizations of differences in f0.

often confronted. Each of the four analytic decisions in 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 f0 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-C illustrate, we could measure f0 at specific points in time like the onset of the temporal window, the offset, or the midpoint. We could also measure the value or time of the f0 minimum or maximum. We could summarize f0 across the entire window and extract the mean, median or standard deviation of f0, all of which have been used to analyze speech data in previous work (see Gordon and 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 f0 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 are multiplied by those researcher degrees of freedom related to statistical analysis (e.g. Wicherts et al. 2016).

In sum, speech data is made of complex physical signals that generate an as-of-yet unappreciated amount of analytic flexibility in the choice of measures and operationalizations. The present paper 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.* In the following section we introduce the research question and the experimental procedure of said project, and we describe the resulting data set as used in the current study.

*Results of this research project were neither published nor publicly presented and are stored on a private OSF repository.

The data set: The 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 non-fruits, other fruits, or other bananas) plays a large part in determining the choice of referring expressions. Generally, speakers 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 over-informative. Speakers use referring expressions that are more specific than strictly necessary for the unambiguous identification of the intended referent (Sedivy 2003; Rubio-Fernández 2016), 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 to 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 over-informative 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. 2013), i.e. what words to use.

Even when morphosyntactically identical expressions are involved, speakers can modulate utterances via acoustic properties like 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 suggest 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 in order to facilitate listeners' comprehension. However, the above studies are based on a small sample size (10 participants) and on potentially anti-conservative 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, thirty 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 the opposite sides of a table, each facing one of two computer screens. The participant and the experimenter could not see each other nor 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 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, which could be moved by the experimenter. Participant 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.

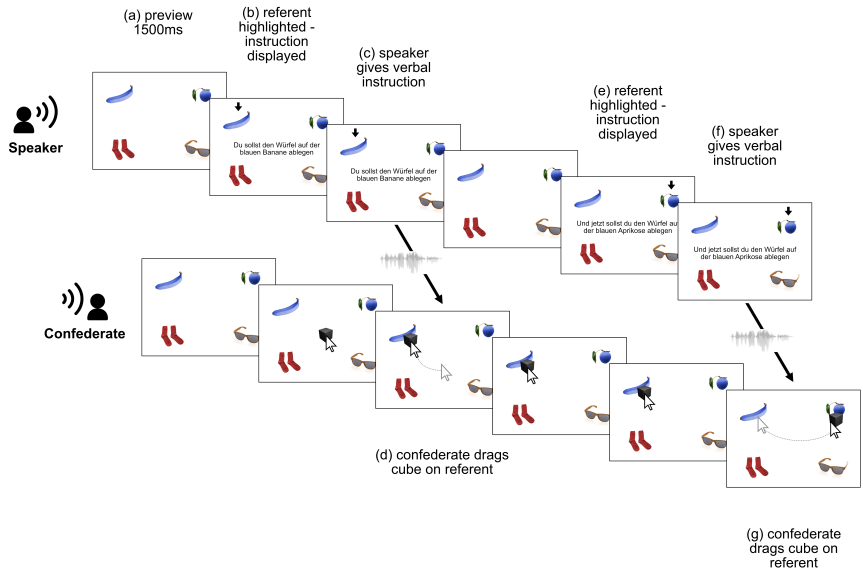

Figure 2. Experimental procedure. The upper row illustrates the trial sequence for the speaker (participant) and the lower row illustrates the trial sequence for the confederate. After a preview of 1500ms the speaker sees an arrow indicating one of the referents (b). Reading the orthographic instructions 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) which the confederate follows by dragging the cube onto the next referent (g).

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 (Mati'c and Wedgwood 2013; Burdin et al. 2015). 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 condition, NF). 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 condition, AF). 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 condition, ANF). The NF condition constituted the experimentally relevant condition, while the AF and ANF conditions acted as fillers. Crucially, the color-object combinations in the Noun Focus (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.[†] 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 present study, 46 analysis teams have 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?*

Methods

As outlined in Section Operationalizing speech, 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 said measurements before submitting them to statistical models, which must be chosen alongside inferential criteria. The complexity of speech data constitutes the ideal testing ground to assess the upper bound of analytic flexibility that social science might face across disciplines. We employed a meta-analytic approach to assess (i) the variability of the reported effects, and (ii) how analytic and researcher-related predictors affect the final results.

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:

[†]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`.

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 nine months. Phase 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 paper and can be accessed at `https://osf.io/3bmcp/`. The repository holds three main OSF components (`Data`, `Teams analyses` and `Questionnaires`), and a link to the project's GitHub repository. 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 pre-recorded 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 (`recruitment_form.pdf`). 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 (`analytic_approach_quest.pdf`). Researchers could choose to work independently or in small teams. For the sake of simplicity, we

will refer both to a single researcher and teams as ANALYSIS TEAMS.[‡] 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`, Section 2.[§] Variability among teams was operationalized as the standard deviation of the teams' reported effects from Silberzahn et al. (2018) (which we z-scored prior to simulations to make it comparable to our study). For the mean of the teams' true standard deviation (0.68 z-score), 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 in order 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 have sampled as many analysts as possible. We received initial expressions of interest to participate from more than 200 analysts, though there was a substantial drop-out rate (see Section Results).

After submitting their analyses, we asked the analysts to also function as peer-reviewers. Each team had to review four other analyses. All analysts involved share co-authorship on this manuscript 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 (`intake_form.pdf`). A PDF version of this and all other questionnaires are available in the repository's `Questionnaires` component, at

[‡]Terms in small caps in this and later sections are included with their definition in the glossary at the end of the paper for the reader's convenience.

[§]Cached model outputs can be found at `https://osf.io/wds2m/`.

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 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 Section The data set. Data included the audio recordings with corresponding time-aligned transcriptions in the form of Praat TextGrid files. These can be found in the Data component at https://osf.io/5agn9/.

Once their analysis was complete, they answered a structured questionnaire (analytic_quest.pdf), providing 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 pre-process 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 is 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 analysis code and the data set. The reviewer was then asked in an online questionnaire (peer_review_quest.pdf) to rate both the acoustic and the statistical analyses and to provide an overall rating, using a scale of 0-100, respectively. To help reviewers calibrate their rating, they were given the following guidelines:

- 100. A perfect analysis with no conceivable improvements from the reviewer.

- 75. An imperfect analysis but the needed changes are unlikely to dramatically alter the final interpretation.
- 50. A flawed analysis likely to produce either an unreliable estimate of the relationship or an over-precise estimate of uncertainty.
- 25. A flawed analysis likely to produce an unreliable estimate of the relationship and an over-precise estimate of uncertainty.
- 0. A dangerously misleading analysis, certain to produce both an estimate that is wrong and a substantially over-precise 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.[¶]

Phase 4: Evaluating variation

The initiating authors (SC, JC, TR) 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. f0 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 *Descriptive statistics* of the Results section.

[¶]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 due to time constraints.

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 below, we have also run further non-preregistered analyses. For these we refer the reader to that section, while we only describe the preregistered analyses in the following paragraphs.

Based on 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) suggest that directly asking analysts to report standardized effect sizes could bias the choice of analyses towards types that more straightforwardly return a standardized effect. Since 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 have 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 (Team 2021), RStan (Team 2020), and brms (Bürkner 2017) in R (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`.

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 multi-level (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_t} ($HalfCauchy(0, 1)$). We fit this model with 4 chains of Hamiltonian Monte-Carlo sampling for the estimation of the joint posterior distribution, using the No U-Turn Sampler (NUTS) as implemented in Stan (Team 2021), and 4000 iterations (2000 for warm-up) per chain, distributed across 8 processing cores and 2 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, detailed in the next section. Crucially, the posterior distribution of the group-level standard deviation σ_{α_t} (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 deemphesizes 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:

$$\text{SDI} = \frac{2|X \cap Y|}{|X| + |Y|}$$

where $|X \cap Y|$ 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. In order 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*, *f0*, *other spectral properties* (e.g. frequency, center of gravity, harmonics

difference, etc.), and *other measures* (e.g. derived measures such principal components, vowel dispersion, etc.). 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*. Since 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 will 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 based on the 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, this 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 mean 0 and standard deviation 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 the Open Science Framework (`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.

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 does not produce an outcome that could be interpreted as an answer to our primary question. For a list of exlcusion criteria, see Section Descriptive statistics 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.[‖]

[‖]The comment history can be found at `https://docs.google.com/document/d/1CFgRo93mRgifpuFOuQE3vNBeMW-H7ps9eD--vxH-6CQ/edit?usp=sharing`.

Results

The results 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 in the GitHub repository linked in the research compendium at `https://osf.io/3bmcp/`, in the `scripts/r/` folder.

Descriptive statistics

In the following sections, we will 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 analysis submitted by the 46 teams, the initiating authors identified analyses from 30 teams to be eligible to be included in the meta-analytic model. Reasons for exclusion were: use of Generalized Additive Models (4 teams) which do not lend themselves easily to the meta-analytic methods employed in this study, use of machine learning techniques (3 teams), use of typicality as the outcome variable/response (3 teams), or use of other methods that returned statistics that could not be included in the meta-analytic model.

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 in the supplementary materials.

The included analyses were provided by 30 teams, comprising 109 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, prior to the data being

released to the teams. Analysts had a median of 5.8 years of experience after completing their PhD, ranging from -3.8 years, i.e. 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 48.5 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 grand total of 104 individual models (including 170 critical model coefficients, given that some models returned more than one critical coefficient) to answer the research question, with a median of 4 models per team. Table 1 provides a summary of the contributing teams and their analyses.

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-six percent of the models used f0 as the outcome variable, 35% used a measure of duration, 12% used vowel formants, 14% 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), 43% 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"), 4% from the whole sentence, and 4% used a different time window. Based on a coarse coding of how acoustic measures were operationalized, we find a total of 52 different measurement specifications. For example, if we consider those analyses that target f0, we find that it is operationalized in many different ways including the minimum, the maximum, the mean, the median, as a range in an interval or a ratio between two intervals. The measurement is sometimes taken from the interval of a vowel in the article, the adjective or noun; it is sometimes taken from the word interval of the article, adjective or noun; or it is 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 = `round(predictor_n$min_predictor,1)-round(predictor_n$max_predictor,1)`). 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, 72% of models contained typicality as a categorical variable (e.g. atypical vs. typical), 24% used a continuous typicality scale from 0-100 by calculating the mean typicality for each word combination as obtained from the norming study, while 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 analysts' strong preference.

The majority of models were run within a frequentist framework (83%). Seventeen percent were run within a Bayesian framework. While 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 contains 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 non-independence is by using models that include so-called random or group level effects (e.g., Gelman and Hill 2006; Schielzeth and Forstmeier 2009), variably known as mixed-effects, hierarchical, multi-level, or nested models (among other names). Eight percent of the linear models specified no random effects at all

(without pooling their data), effectively ignoring these non-independences (Hurlbert 1984). Sixty-five percent specified random intercepts only, and 27% specified both random intercepts and random slopes to account for the non-independence. 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 47 different model specifications.

When considering both acoustic and statistical analyses, we have found a grand total of 104 different analytic pipelines. In other words, each individual analysis submitted was unique.

Our quantitative assessment did not include other degrees of freedom, all of which are additional sources of variation: Teams differed 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; teams differed in whether the statistical analysis was based on a subset of the data or the whole dataset; and they differed whether and if so how measurements were excluded based on 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. Nine teams out of the thirty (30%) reported to have found at least one statistically reliable effect (based on the inferential criteria they specified). Of the 104 submitted models, 37 were reported to show a statistically reliable effect (21.8%).

Review ratings Teams reviewed each others' acoustic and statistical analyses. The mean rating of the acoustic analyses, on a scale from 0 to 100, is 71.5 (SD = 13.9). The mean rating of the statistical analysis is 69.5 (SD = 16.5). For reference, as mentioned in the Methods 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".

Table 1. Descriptive statistics of teams, acoustic analyses, and statistical analyses included in the meta-analysis. The data set included analyses from 30 teams and 109 analysts.

Team characteristics		**Range**	**Median**
	Team size	1.0 – 12.0	3.0
	Years after PhD	-3.8 – 12.4	5.8
	Prior belief	48.5 – 92.0	70.0
	Acoustic analysis peer rating	41.2 – 88.3	75.2
	Statistical analysis peer rating	33.0 – 93.3	74.1
	Overall peer rating	39.0 – 88.7	72.2
Acoustic analyses		**n**	**%**
Outcome	F0	37	36
	Duration	36	35
	Intensity	15	14
	Formants	13	12
	Other	3	3
Temporal window	Segment	47	46
	Word	45	44
	Sentence	4	4
	Phrase	3	3
	Other	4	4
Typicality operationalization	Categorical	75	73
	Continuous (mean)	25	24
	Continuous (median)	3	3
Statistical analyses		**n**	**%**
Framework	Frequentist	86	83
	Bayesian	18	17
Model	Linear model	102	98
	GAM	1	1
	Other	1	1
		Range	**Median**
N	Models	1 – 16	4
	Predictors	1 – 5	2
	Random terms	1 – 10	2
	Intercept	1 – 10	2
	Slope	0 – 4	1

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

Between-team variability The primary aim of this analysis is 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 is 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 ± 0.06 (0.07 * 1.96) with 95% credibility."

Non-preregistered. However, in our preregistration we did not take into account that teams might submit multiple analyses/models which, if unaccounted for, violates the independence assumption. Teams were explicitly instructed to only submit one effect size without enforcing it. As a result, some teams followed the instruction and submitted only one model while others submitted multiple models. To account for this added layer of dependency, we have run a model with team and model ID nested within team as group-level effects (`(1|team) + (1|team:model_id)`), which allows 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.**

The nested model yields a posterior 95% CrI for between-team variability of 0 to 0.05 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 0.15 standard units ($\beta = 0.132$, SD = 0.01). For the sake of illustration, these would correspond to an estimate of between-model variability in milliseconds (if looking for example at segment

**Note that before fitting this model, we fit a separate one in which model ID was the only (non-nested) group-level effect. The estimated group-level effect of model ID is identical to that of the nested model, so we will not discuss it further.

duration) that ranges between 7.4 and 15.1 ms at 95% credibility. We interpret these values in more details 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 assessing 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 -0.024 and 0.021 standard units (95% CrI, mean = -0.001). In other words, our best guess is that speakers might not encode typicality in the acoustic signal (e.g. by duration, f0, etc,) or, if they do, they do so by a maximum of ± 0.02 standard units.

Non-preregistered. As mentioned in the previous section, we have run 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 is between -0.013 and 0.04 standard units (95% CrI, β = 0.014). This suggests that the acoustic measures of typical word combinations are 0.01 standard units lower to 0.04 standard units higher than the measures of atypical word combinations, at 95% confidence. Relative to the preregistered model, this model suggests a somewhat positive effect of typicality (although small negative effects are also possible).

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, etc). This means that some of the effects on a particular response as observed in a specific location within the utterance might naturally be positive, while other 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. More on this 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

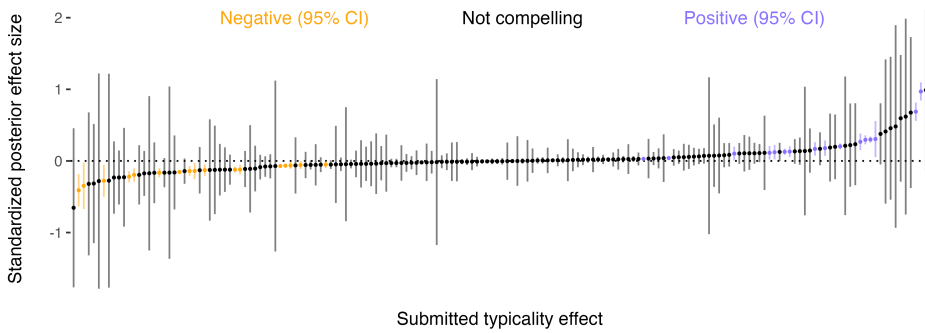

Figure 3. Standardized effect sizes across all critical coefficients provided by the teams. Raw estimates are displayed in grey. Estimates after shrinkage as provided by the meta-analytic model are displayed in black/orange.

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 (grey triangles and lines), it is striking how much estimates differed. Estimates ranged from -0.65 to 0.99 standard units.

While the majority of model estimates and their uncertainty after shrinkage (in black) yields inconclusive results (i.e. are compatible with a point null hypothesis), there are 27 model estimates for which the 95% credible interval does not contain zero (in blue, 16%).

Analytic and researcher-related predictors

After assessing 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).

We have run a model as described in Section Analytic and researcher-related predictors affecting effect sizes above. Figure 4, panel C, displays the coefficients for all predictors alongside their 80% and 95% credible intervals. The model suggests that most team-specific predictors yield very small deviations from the meta-analytic estimate and their 95% credible intervals include zero, leaving us highly uncertain about their direction. Neither analysts' prior beliefs in the phenomenon (β = 0.00, 95% CrI = [-0.03, 0.03]), nor their seniority in terms of years after completing

their PhD (β = 0.01, 95% CrI = [-0.03, 0.04]) seem to affect model estimates. Similarly, the evaluation of the quality of the analysis from their peers yielded a rather small effect magnitude, again characterized by large uncertainty (β = 0.01, 95% CrI = [-0.02, 0.05]). Interestingly, the model uniqueness, i.e. how unique the choice and combination of predictors are, affects the analysts' estimate, with more unique models producing higher positive estimates (β = 0.05, 95% CrI = [0.02, 0.07]).

Looking at the most important choices during measurement, both the acoustic parameter under investigation (e.g. f0 or duration) and the choice of measurement window affected the results. Panels A and B of Figure 4 display the posterior estimates for the measurement outcome (i.e. what acoustic dimension was measured, panel A) and measurement window (i.e. what is the unit over which the outcome was measured, panel B). If, on one hand, an acoustic dimension related to f0 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 of the latter patterns that we need to interpret the results of the model with great caution. Since 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 while 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*

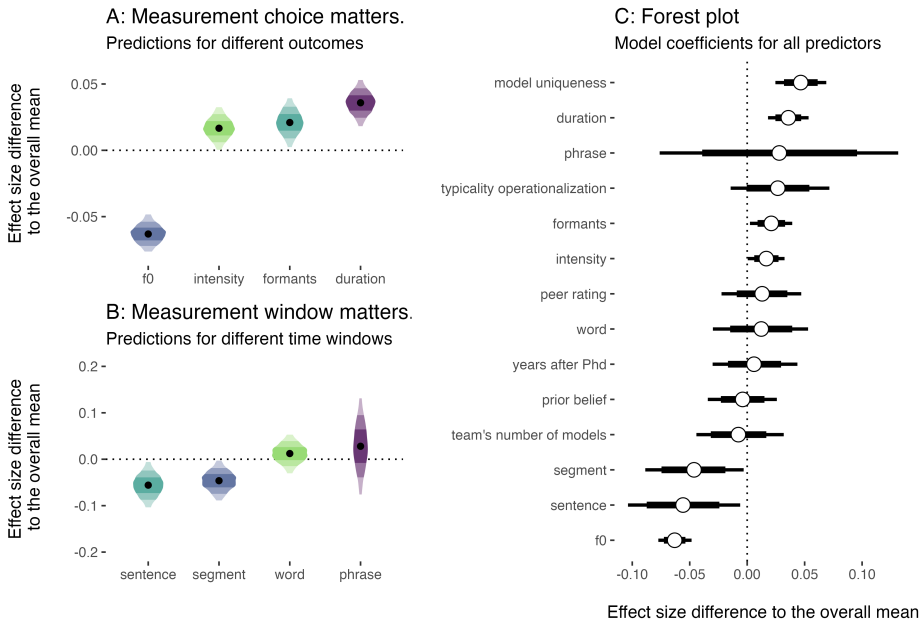

Figure 4. The effects of analytic and researcher-related predictors on the reported standardized effect sizes. (A) Posterior samples for the four most frequent outcome variables; (B) Posterior samples for the four most frequent temporal windows: Black points indicate medians; shaded areas represents 50/80/95% highest density intervals. (C) Mean posterior samples (white circles) and 80/95% credible intervals for all predictors.

atypical word combinations? In order to answer this question, teams had to interpret the research question by operationalizing constructs 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", i.e. to a wide range of combinations of possible analytic decisions. The submitted analyses exhibited at least 47 unique ways of operationalizing the acoustic signal alongside 52 unique ways of constructing the statistical model. By multiplying the numbers of acoustic and model specifications, there are in principle 2444 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 like e.g. 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 30% 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, f0 and intensity, taken from different time windows. These standard deviations can be considered as 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, f0 and intensity.[††]

For example, for those analyses that investigated the duration of vowels (e.g. the duration of the stressed vowel in *Banane*), 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 units. The reported effect estimates from teams varied between -0.65 and 0.99 standard units, which corresponds to duration values ranging from -21.86 to 50.81 ms. A more conservative approach is to convert the meta-analytic estimates of between-model variation, thus obtaining an estimate of between-model variability in milliseconds that ranges between 7.4 and 15.1 ms at 95% credibility.[‡‡]

While 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 and 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

[††]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 their metrical structure.

[‡‡]The calculation is thus: the minimum standard deviation of duration multiplied by the lower limit of the 95% CrI of the between-models variability estimate, times 1.96 to obtain a 95% CrI: $33.4 * 0.11 = 3.8$, $3.8 * 1.96 = 7.4$ ms; the maximum standard deviation of duration multiplied by the upper limit of the 95% CrI of the between-models variability estimate, times 1.96: $51.4 * 0.15 = 7.7$, $7.7 * 1.96 = 15.1$ ms.

Table 2. Estimated 95 percent CrIs 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	7.4-11.4	10-15.4	ms
Duration	Word	7.2-26.5	9.8-35.8	ms
f0	Segment	8.9-9.8	12-13.3	hz
f0	Word	0.9-10.4	1.2-14.1	hz
Intensity	Segment	0.7-1.6	1-2.1	dB
Intensity	Word	0.8-0.9	1-1.2	dB

with segmental contrasts (Coretta 2019), reduction phenomena (Nowak 2006), and biomechanical reflexes of prominence (Mücke and Grice 2014). Thus, variation between different analyst teams of 7 to 15 ms in one or the other direction can be theoretically relevant and might lead to opposing theoretical conclusions.

While one might find it obvious that measuring different parts of the speech signal can lead to different results, the fact that analysts (and reviewer 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 choose to measure more or less the same acoustic property within the same measurement window, they arrive at different estimates: For example, six teams measured f0 in the adjective and predicted f0 based on typicality as a categorical predictor. Their standardized effect estimates ranged from -0.11 to 0.38 standard deviations. While these teams in principle measured the same thing, they differed in analytical details of how f0 was operationalized (i.e. mean, mininum, maximum, or range) and how their statistical model was constructed (i.e. the number of predictors ranged from 1-3 and the number of random effect terms ranged from 1-4). As shown by (2021), even seemingly inconsequential analytical choices can affect conclusions in non-trivial 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 crowd-sourced projects that suggest variation between teams is neither driven by perceived quality of the analysis nor by analysts' biases or experience (e.g., Silberzahn et al. 2018; Breznau et al. 2021). Following Breznau et al. (2021, p. 9), 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". Idiosyncratic variation across researchers might be a fact of life which we have to acknowledge and integrate into how we evaluate and present evidence.

While 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 while others seem to result in estimates into another. For example, measurements related to f0 tended to result in lower estimates while 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.

Secondly 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 f0, likely because similar functional relationships like information structure and predictability can be expressed by f0 (e.g. Grice et al. 2017; Turnbull 2017). Moreover, prior work has actually suggested a relationship between typicality and f0 (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 into 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 employed. Rather, we simply want to point out that if the theoretical underpinnings of the field were much clearer, different teams would have converged towards a limited set of analyses despite of 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 statement 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. While 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 observed in the present study can not be corrected for in a straightforward way. If uncorrected for, 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 (Kerr 1998; John et al. 2012; Simmons et al. 2011), while null results remain unreported (Sterling 1959; Rosenthal 1979). 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, i.e. 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 30% 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 -29, 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 (Vazire 2017; Fidler and Wilcox 2018). One of the suggested paths forward is for researchers to directly replicate previous studies more often (Open Science Collaboration 2015; Camerer et al. 2018). While we agree with the importance of direct replications, our study (and similar crowd-sourced 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. Beyond a replication

crisis, quantitative disciplines are going through what has been called 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 (see Campbell 1975, for an overview of the concept).

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 regards 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)—prior to 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.

In light of the observed analytic flexibility, there are a few things that researchers can do to appropriately calibrate confidence in their claims. First of all, 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 and Weenink 2021; EMU: Winkelmann et al. 2017; the Montreal Forced Aligner: McAuliffe et al. 2017).

Second, making analytic pathways completely re-traceable does not change the fact that analysts apply different analytic approaches. Crowd-sourced 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. Crowd-sourcing 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.

Third, 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, Steegen et al. 2016; Harder 2020) and has recently gained popularity across disciplines.

Finally, neither crowd-sourcing nor multiverse analyses will guarantee that all relevant pathways are explored. Crowd-sourcing 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, i.e. in the present case how communicative

pressures of certain functions are expressed in the acoustic signal. Possible tools to strengthen theoretical development are relate to mathematically formalizing verbal expectations or using computational models (e.g., van Rooij and Blokpoel 2020; Guest and Martin 2021; Scheel et al. 2021; Devezer et al. 2021).

Caveats

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

First, while the total number of analyses is larger than most earlier crowd-sourcing projects, it is likely to be too small to reliably estimate the impact of certain predictors. Since 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 and 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. Do speakers modify f0 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 and Roettger 2017). Research on acoustic markers of clinical conditions such as depression and schizophrenia are often difficult to compare due to the wide variety of different acoustic measures employed (e.g. Cummins et al. 2015; Parola et al. 2022).

Third, the design of this crowd-sourced study has artificially inflated the variability between teams by encouraging anti-coordination strategies.

Teams knew that there will other analyst teams and therefore might have chosen a "less canonical" analysis. Since analysts were guaranteed to become co-authors of a (in principle) guaranteed publication, such an anti-coordination approach was not explicitly disincentivized.

Forth, our sample is an opportunity sample. We have advertised the project through online platforms which 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., Jafar 2018; Darwin Holmes 2020; Fox et al. 2021). 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 judgement 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 re-analyze our data and try to replicate our results using a different research question.

Finally, the present 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 sub-disciplines 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). While we do believe that our qualitative conclusions holds 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 and Van Grootel 2019). It is conceivable that the discussed issues apply differently or not at all to qualitative data analyses.

Conclusion

In recent efforts, several 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. is there evidence for a relationship or not) and different effect magnitudes. This is concerning, as 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, what we find of particular relevance is the fact that commonly research proceeds based on publications from one research team at a time. If we imagine a situation where 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 like the one employed 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 paper 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 bio-mechanical 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 inconclusive results. We suggest that fields that use multidimensional data need to acknowledge these degrees of freedom, consider crowd-sourcing 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.

Author contributions

See `https://github.com/many-speech-analyses/many_analyses/blob/main/figs/credit-taxonomy.png`.

Conflicts of interest

We have no conflicts of interest to disclose.

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.

References

Aczel B, Szaszi B, Nilsonne G, Van den Akker O, Albers CJ, van Assen MALM, Bastiaansen JA, Benjamin DJ, Boehm U, Botvinik-Nezer R and et al (2021) Guidance for multi-analyst studies. DOI:10.31222/osf.io/5ecnh.

Arts A, Maes A, Noordman LG and Jansen C (2011) Overspecification in written instruction. *Linguistics* 49(3): 555–574.

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

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

Bastiaansen JA, Kunkels YK, Blaauw FJ, Boker SM, Ceulemans E, Chen M, Chow SM, de Jonge P, Emerencia AC, Epskamp S et al. (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: 110211.

Boersma P and Weenink D (2021) Praat: doing phonetics by computer [computer program](2011).

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA et al. (2020) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582(7810): 84–88.

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

Breznau N, Rinke EM, Wuttke A, Adem M, Adriaans J, Alvarez-Benjumea A, Andersen HK, Auer D, Azevedo F, Bahnsen O et al. (2021) Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty DOI:10.31222/osf.io/cd5j9.

Brugger P (2001) From haunted brain to haunted science: A cognitive neuroscience view of paranormal and pseudoscientific thought. *Hauntings and Poltergeists: Multidisciplinary Perspectives* : 195–213.

Burdin RS, Phillips-Bourass S, Turnbull R, Yasavul M, Clopper CG and Tonhauser J (2015) Variation in the prosody of focus in head- and head/edge-prominence languages. *Lingua* 165: 254–276. DOI:https://doi.org/10.1016/j.lingua.2014.10.001. URL `https://www.sciencedirect.com/science/article/pii/S0024384114002290`. Prosody and Information Status in Typological Perspective.

Bürkner PC (2017) brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1): 1–28. DOI:10.18637/jss.v080.i01.

Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T et al. (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour* 2(9): 637–644. DOI: 10.1038/s41562-018-0399-z.

Campbell DT (1975) On the conflicts between biological and social evolution and between psychology and moral tradition. *American psychologist* 30(12): 1103.

Charles SJ, Bartlett JE, Messick KJ, Coleman TJ and 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 and Quatieri TF (2015) A review of depression and suicide risk assessment using speech analysis. *Speech communication* 71: 10–49.

Darwin Holmes AG (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. DOI:10.34293/education.v8i4.3232.

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

Degen J, Hawkins RD, Graf C, Kreiss E and Goodman ND (2020) When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review* .

Devezer B, Navarro DJ, Vandekerckhove J and Ozge Buzbas E (2021) The case for formal methodology in scientific reform. *Royal Society open science* 8(3): 200805.

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

Dimitrova DV, Redeker G, Egg M and Hoeks JC (2008) Prosodic correlates of linguistic and extra-linguistic information in dutch. In: *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, pp. 2191–2196.

Dimitrova DV, Redeker G and Hoeks JC (2009) Did you say a blue banana? the prosody of contrast and abnormality in bulgarian and dutch. In: *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).

Dutilh G, Annis J, Brown SD, Cassey P, Evans NJ, Grasman R, Hawkins GE, Heathcote A, Holmes WR, Krypotos AM et al. (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 and Wilcox J (2018) Reproducibility of scientific results URL `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.

Flake JK and Fried EI (2020) Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science* 3(4): 456–465. DOI:10.1177/2515245920952393.

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

Fox J, Pearce KE, Massanari AL, Riles JM, Szulc Ł, Ranjit YS, Trevisan F, Soriano CRR, Vitak J, Arora P et al. (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 RP, van Deemter K and Krahmer E (2013) Are we bayesian referring expression generators. In: *Proceedings of the CogSci workshop on the production of referring expressions*. pp. 1–6.

Gelman A and Hill J (2006) *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

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

Grice HP (1975) Logic and conversation. In: *Speech acts*. Brill, pp. 41–58.

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

Guest O and Martin AE (2021) How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science* 16(4): 789–802.

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

Haven TL and Van Grootel DL (2019) Preregistering qualitative research. *Accountability in research* 26(3): 229–244.

Hawkins S and 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 SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecological monographs* 54(2): 187–211.

Jafar AJN (2018) What is positionality and should it be expressed in quantitative studies? *Emergency Medicine Journal* DOI:10.1136/emermed-2017-207158.

John LK, Loewenstein G and 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 and Wong S (2000) Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America* 108(3): 1252–1263.

Kerr NL (1998) Harking: Hypothesizing after the results are known. *Personality and social psychology review* 2(3): 196–217.

Klein O, Hardwicke TE, Aust F, Breuer J, Danielsson H, Mohr AH, IJzerman H, Nilsonne G, Vanpaemel W, Frank MC et al. (2018) A practical guide for transparency in psychological

science. *Collabra: Psychology* 4(1).

Knight K (2000) *Mathematical Statistics*. Chapman and Hall, New York.

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

Ladd DR (2008) *Intonational phonology*. Cambridge University Press.

Landy JF, Jia ML, Ding IL, Viganola D, Tierney W, Dreber A, Johannesson M, Pfeiffer T, Ebersole CR, Gronau QF et al. (2020) Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin* 146(5): 451.

LeBel EP, McCarthy RJ, Earp BD, Elson M and 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 and 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 and Sonderegger M (2017) Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In: *Proc. Interspeech 2017*. pp. 498–502. DOI:10.21437/Interspeech.2017-1386.

Meehl PE (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 KM, Gerber A, Glennerster R, Green DP, Humphreys M, Imbens G et al. (2014) Promoting transparency in social science research. *Science* 343(6166): 30–31.

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

Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers EJ, Ware JJ and Ioannidis J (2017) A manifesto for reproducible science. *Nature human behaviour* 1(1): 1–9.

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

Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2): 175–220.

Niebuhr O, d'Imperio M, Fivela BG and Cangemi F (2011) Are there "shapers" and "aligners"? individual differences in signalling pitch accent category. In: *Proceedings of the 17th International Congress of Phonetic Sciences*. pp. 120–123.

Nosek BA and 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 Thesis.

Ogden R (2004) Non-modal voice quality and turn-taking in Finnish. *Sound patterns in interaction: cross-linguistic studies from conversation* : 29–62.

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

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

Parker T, Fraser H, Nakagawa S, Gould EB, Griffith S, Vesk P and Fidler F (2020) Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology [passed peer review and granted in-principle acceptance March 2020]. *BMC Biology* DOI:10.6084/m9.figshare.12034833.v1.

Parola A, Lin JM, Simonsen A, Bliksted V, Zhou Y, Wang H, Inoue L, Koelkebeck K and Fusaroli R (2022) Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of nlp automated measures of coherence. *Schizophrenia Research* .

Port RF and Leary AP (2005) Against formal phonology. *Language* 81(4): 927–964.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Roettger TB (2019) Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1).

Roettger TB, Winter B and 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.

Rotello CM, Heit E and 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: 153.

Scheel AM (2022) Why most psychological research findings are not even wrong. *Infant and Child Development* 31(1): e2295.

Scheel AM, Tiokhin L, Isager PM and Lakens D (2021) Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science* 16(4): 744–755.

Schielzeth H and Forstmeier W (2009) Conclusions beyond support: overconfident estimates in mixed models. *Behavioral ecology* 20(2): 416–420.

Sedivy JC (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 EL, Martin DP, Anselmi P, Aust F, Awtrey E, Bahník Š, Bai F, Bannard C, Bonnier E et al. (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.

Simmons JP, Nelson LD and 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.

Starns JJ, Cataldo AM, Rotello CM, Annis J, Aschenbrenner A, Bröder A, Cox G, Criss A, Curl RA, Dobbins IG et al. (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.

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

Sterling TD (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 KN (2000) *Acoustic phonetics*, volume 30. MIT press. DOI: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.

Team SD (2020) RStan: the R interface to Stan. URL `http://mc-stan.org/`. R package version 2.21.2.

Team SD (2021) Stan modeling language users guide and reference manual, v2.26.0. URL `http://mc-stan.org/`.

Tukey JW (1977) *Exploratory data analysis*, volume 2. Reading, MA.

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

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

Van Heuven VJ, Haan J, Gussenhoven C and Warner N (2002) Temporal distribution of interrogativity markers in Dutch: A perceptual study. In: *Laboratory Phonology*, volume 7. Walter de Gruyter, pp. 61–86.

Van Heuven VJ and 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 and Blokpoel M (2020) Formalizing verbal theories: A tutorial by dialogue. *Social Psychology* 51(5): 285.

Vazire S (2017) Quality uncertainty erodes trust in science. *Collabra: Psychology* 3(1).

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL and Kievit RA (2012) An agenda for purely confirmatory research. *Perspectives on Psychological Science* 7(6): 632–638.

Westerbeek H, Koolen R and Maes A (2015) Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in psychology* 6: 935.

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

Wicherts JM, Borsboom D, Kats J and Molenaar D (2006) The poor availability of psychological research data for reanalysis. *American psychologist* 61(7): 726.

Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM and van Assen MALM (2016) Degrees of freedom in planning, running, analyzing, and reporting psychological

studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7. DOI:10.3389/fpsyg.2016.01832.

Winkelmann R, Harrington J and 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.