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

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

Abstract

Recent efforts to replicate published findings have uncovered surprisingly low success rates across disciplines. Moreover, several 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 that derives not only from aspects of statistical modelling but also from decisions regarding the quantification of the measured behavior. In the present study, we will give the same speech production dataset to N teams of researchers and will ask them to answer the same research question. Using Bayesian meta-analytic tools, we will quantify this variability and relate it to predictors related to the analysts and their analyses.

Keywords

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

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., 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 the 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.

[1]Institute of Phonetics and Speech Processing, Ludwig-Maximilian University Munich, Germany.
[2]Department of Spanish and Portuguese, Rutgers University, New Brunswick, United States
[3]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

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 human speech production, for example, researchers need to make numerous decisions about what to measure and how to measure it, in other words, how to operationalise 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 analyze the same speech production data set, a process that involves both decisions regarding the *operationalization* of a complex signal and decisions regarding *statistical analysis*. Specifically, we discuss the degree of variability in research results obtained by N teams who had to choose the operationalisation 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 operationalisation of complex human behavior (here, human speech). This is an important addition in that the increased number of "forking paths" in the "garden of analytic choices", deriving from the many decisions involved in quantification, might reveal a higher degree of variability across analysts than what previously observed, thus giving us a more realistic estimate of variability across analysts.

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 filter the world in irrational ways (e.g., Tversky and Kahneman 1974), 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, it is likely that a good part of published papers 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).

Crowdsourcing alternative analyses

Recent collaborative attempts have started to shed light on how different analysts tackle the same data set and have revealed a large amount of

variability. In a seminal collaborative effort, Silberzahn et al. (2018) let twenty-nine independent analysis teams address the same research hypothesis. Analytic approaches and consequently the results varied widely between teams. Sixty-nine percent of the teams found support for the hypothesis, and 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 team's 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) on analytic flexibility in ecology, and Botvinik-Nezer et al. (2020) on neuroimaging data). Bastiaansen et al. (2020) crowdsourced clinical recommendations based on analyses of an individual patient. Their results suggests that analysts differed substantially regarding decisions related to both the statistical analysis of the data and the theoretical rationale behind interpreting the statistical results.

Extending on the many-analysts approach, Landy et al. (2020) asked 15 research teams to independently design studies to answer the same research questions. Again, they found vast heterogeneity in researchers' conclusions which was not predicted by the researchers' expertise but seem to be more dependent on the research question itself. This is in line with a recent re-analysis of Silberzahn et al. (2018) by Auspurg and Brüderl (2021). The authors argue that 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 statistical inference 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 path related to measurement.

However, in many fields the primary raw data are complex signals that 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 a theoretical construct related to human behavior or its underlying cognitive processes 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, and 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 therapy 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 is a complex signal that is characterized by structurally different acoustic landmarks distributed throughout different temporal domains. Thus, choosing how to measure a communictive intention of interest is an important analytic step. Take for example the following 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 classical 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 position of the consonant in the word and 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 for 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 pitch 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 with 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

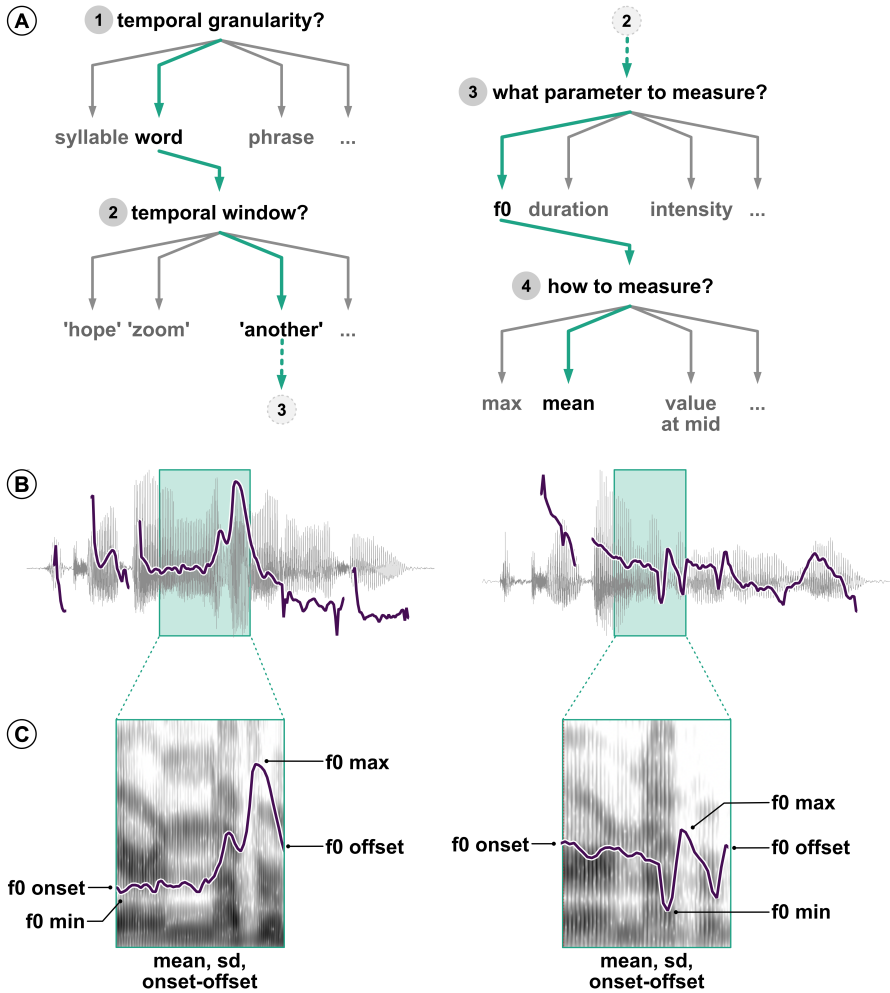

Figure 1. Illustration of the analytic flexibility associated with acoustic analyses. (A) An example of multiple possible and justifiable decisions when comparing two utterances; (B) Waveform and fundamental frequency (f0) track of the utterances *I can't bear ANOTHER meeting on Zoom* and *I can't bear 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.

or do we measure a specific point in time? Do we measure fundamental frequency, intensity, or something else (Stevens 2000)?

When looking at phrase-level temporal domains, the number of alternative analytic pipelines increases substantially. Figure 1A shows a typical example of a decision tree with which speech researchers are often confronted. Each of the four analytic decisions in 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 fundamental frequency 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 minimum or maximum f0 value. 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). These myriads 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 generates 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

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

and the experimental procedure of said project, and we describe the resulting data set as used in the current study.

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 a yellow and a less ripe green banana available, but not when there is only one banana to begin with.

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; Westerbeek et al. 2015; 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 a referring expression depends on how good it is 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). 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 suprasegmental 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 suggested that speakers might use intonation to signal atypical lexical combinations (e.g. Dimitrova et al. 2008, 2009). Referential expressions such as *blue banana* were produced with greater prosodic prominence than more typical referents such as *yellow banana*. These results are in line with the idea of resource-efficient, rational language users who modulate their speech 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. The participant would 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,

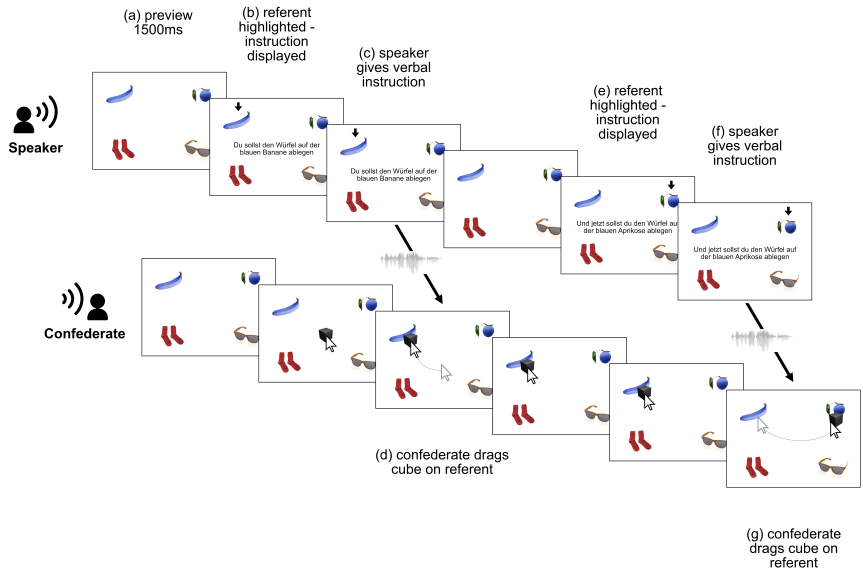

Figure 2. Experimental procedure. The upper row illustrates the trial sequence for the speaker (participant) and the lower row illustrates the trial sequence for the confederate. After a preview of 1500ms the speaker sees an arrow indicating one of the referent (b). Reading the orthographic instructions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

you have to put the cube on top of the COLOR OBJECT') instructing the experimenter to move the cube on top of a different object (the *target*). The second utterance in the trial was the critical trial for analysis.

The two sentence prompts were used to create a focus contrast between the competitor and the target object. Focused units denote the set of all (contextually relevant) alternatives (e.g. Rooth 1992). Concretely, a focus contrast marks one of more elements in a sentence as prominent, by different linguistic means depending on the language. 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. We call this the Noun Focus condition, or NF for short. If the objects are the same but differ in color (e.g. *yellow banana* vs *blue banana*), the color adjective is in

focus (the 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 (the 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, N analysis teams will receive 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 will be 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 amount 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 will employ 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.

[†]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`.

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

1. RECRUITMENT: We will recruit independent groups of researchers to analyze the data and review others' data analyses.
2. TEAM ANALYSIS: We will give researchers access to the speech corpus and let them analyze the data as they see fit.
3. REVIEW: We will ask reviewers to generate peer-review ratings of the analyses based on methods (not results).
4. META-ANALYSIS: We will evaluate variability among the different analyses and how different predictors affect the outcomes.
5. WRITE-UP: We will collaboratively produce the final manuscript.

We estimate that this process, from the time of an in-principle acceptance of this Stage 1 Registered Report to the end of Phase 5, will take nine months. The factor most likely to delay our time-line is the rate of completion of the original set of analyses by independent analysis teams.

The project OSF repository contains all the materials mentioned in this paper and can be accessed at `https://bit.ly/3AqU1ul`. The main repository holds three OSF components, `Data`, `Questionnaires`, and `Cache`, and it is linked to the 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 will provide 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 will be advertised via social media, using mailing lists for linguistic and psychological societies, and via word of mouth. Social media advertising will be accompanied by a short recruitment form (`recruitment_form.pdf`) The target population comprises active speech science researchers with a graduate/doctoral

[‡]"Questions" and "Join Project" buttons will be functioning when recruitment starts. "Join Project" will link to the intake form.

degree (or currently studying for a graduate/doctoral degree) in relevant disciplines. All individuals interested in participating will be asked to complete a questionnaire detailing their familiarity with numerous analytical approaches common in the speech sciences (`analytic_approach_quest.pdf`). Researchers can choose to work independently or in a small team. For the sake of simplicity, we will refer both to a single researcher and teams as ANALYSIS TEAMS.[§] Recruitment for this project will commence upon receiving 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/suppl/simulations.html`, 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 consider a minimum sample size of 10 teams as sufficient. Given the exploratory nature of our study, however, we will sample as many analysts as possible. We have received initial expression of interest to participate from more than 200 analysts.

After submitting their analyses, we will ask analysts to also function as peer-reviewers. Each team must review four other analyses. All analysts will share co-authorship on this manuscript and will participate in the collaborative process of producing the final manuscript. Informed consent will be obtained as part of the intake form.

[§]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://bit.ly/3AgdHAN`.

Phase 2: Primary Data Analyses

The analysis teams will register for participation and each of the analysts individually will answer 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 `https://bit.ly/3Cf3fdm`. The questionnaire collects information on the analysts current position and self-estimated breadth and level of statistical expertise and acoustic analysis skills. We will then request that they answer the question: *Do speakers acoustically modify utterances to signal atypical word combinations?* To do so, they will be given the data generated by the experiment described in Section The data set. The data will include 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://bit.ly/3Ahawc7`.

Once their analysis is complete, they will answer 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 will also upload 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

Each analysis will be evaluated by four different analysts teams who function as peer-reviewers. Each peer-reviewer will be randomly assigned to analyses from at least four analysis teams. Reviewers will evaluate the methods of each of their assigned analyses one at a time in a sequence determined by the initiating authors. The sequences will be 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 will be as follows. First, the reviewer will receive a description of the methods of a single analysis. This will include the narrative methods section, the analysis team's answers to the questionnaire regarding their methods, including analysis code and the data set. The reviewer will then be 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 will be 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 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 will also be 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 will then be made available to the reviewer. This same sequence will be followed until all analyses allocated to a given reviewer have been provided and reviewed. After providing the final review, the reviewer will be simultaneously presented with all four (or more) methods sections that the reviewer has just completed reviewing, the option to revise their original ratings, and a text box to provide an explanation.

Phase 4: Evaluating variation

The initiating authors (SC, JC, TR) will conduct the analyses outlined in this section. We will not conduct confirmatory tests of any a priori hypotheses. We consider our analysis exploratory.

Descriptive statistics We will calculate 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 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).

Meta-analytic estimation We will investigate the variability in REPORTED EFFECT SIZES using Bayesian meta-analytic techniques. As the measure of variability, we will take the meta-analytic GROUP-LEVEL STANDARD DEVIATION (σ_{α_t}, see below), where the groups are the analysis teams.

Based on the common practices currently in place within the field, we anticipate that researchers will use multilevel regression models, thus common measurements of effect size, such as Cohen's d, may be 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 substantially differ in their measurement scales (e.g, Hertz for frequency vs milliseconds for duration), we will standardize 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. Factor-level ordering for each categorical variable will mirror that of the original analyses. Each STANDARDIZED MODEL will be fitted as a Bayesian regression model with Stan (Team 2021), RStan (Team 2020), and brms (Bürkner 2017) in R (R Core Team 2020). For those reported models that were originally fitted within a frequentist approach, uniform distributions will be used as the priors of all parameters (with the restriction that only positive numbers will be included for scale parameters), to approximate the estimation procedure of frequentist models (i.e. only the distribution of the data determines the posterior distribution). If a team has fitted Bayesian models, the same priors as reported by the team will be used in fitting the respective standardized model. Model refitting will also constitute a way of validating the reported analyses, a step recommended by Aczel et al. (2021).

The coefficients of the critical predictors (i.e. critical according to the analysis teams' self-reported inferential criteria) obtained from the standardized models will be used as the STANDARDIZED EFFECT SIZE (η_i)

of each reported model. If multiple predictors within a single analysis have been reported as critical, each will be included in the meta-analytic model (described in details in the next paragraph). Moreover, to account for the differing degree of uncertainty around each standardized effect size, we will use the standard deviation of each standardized effect size as the STANDARDIZED STANDARD ERROR (se_i). This will enable 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 will be 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 of the individual reported model, we will conduct a BAYESIAN RANDOM-EFFECTS META-ANALYSIS using a multilevel (intercept-only) regression model. The outcome variable will be the set of standardized effect sizes η_i. The likelihood of η_i is assumed to correspond to a normal distribution (Knight 2000). The analysis teams will be entered as a group-level effect (i.e., `(1 | team)`, called *random effect* in the frequentist literature). The standard errors se_i will be included as the standard deviation of η_i to fit a measurement-error model, as discussed above. We will use 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 will 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 4 processing cores. In case of divergent transitions, we will increase `adapt_delta`, `tree_depth`, and the number of iterations in this order until we obtain a model fit with no divergent transitions. The code used to run the model can be found at `https://many-speech-analyses.github.io/suppl/simulations.html`.

The posterior distribution of the population-level intercept α will allow us to estimate the range of probable values of the standardized effect size $\hat{\eta}$. The posterior distribution will further allow 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) will allow 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 will investigate the extent to which the individual standardized effect sizes are affected by a series of ANALYTIC AND RESEARCHER-RELATED PREDICTORS.

Analytic predictors. We will estimate 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].
- *Measure of conservativeness* of the model specification, as the number of random/group-level effects included [numeric].
- *Number of post-hoc changes to the acoustic measurements* the teams will report to have carried out [numeric].
- *Number of models* the teams will report 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 will be 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. For our purposes, we will treat predictors as species and individual analyses as sites. For each pair of analyses (X, Y), the SDI will be 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.

In order to generate a unique SDI for each analysis team, we will calculate 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 will be categorized according to the following possible groups: *duration*, *amplitude*, *fundamental frequency*, *other spectral properties* (e.g. frequency, center of gravity, harmonics difference, etc.), and *other measures* (e.g. 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*. Since each analysis will receive more than one peer-review rating, we will calculate the mean rating and its standard deviation for each. These will be entered in the model formula as a measurement-error term (`me(mean, sd)` in brms).

Researcher-related factors. We will also include 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 students [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 will calculate the mean and standard deviation of these predictors for each team. These will be 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).

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

A normal distribution will be used as the likelihood function of $\alpha_{t[i]}$. The mean of $\alpha_{t[i]}$ is modeled on the basis of the overall intercept β

and on the coefficients of each predictor. The numeric predictors will be centered and scaled and the categorical predictors will be sum coded. As the prior for the intercept and the predictors we will use a normal distribution with mean 0 and standard deviation 1. The model will be 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/suppl/simulations.html`.

Data management All relevant data, code, and materials will be publicly archived on the Open Science Framework (`https://osf.io/3bmcp/`, link for peer-review: `https://bit.ly/3AqU1ul`). Archived data will include the original data set distributed to all analysts, any edited versions of the data analyzed by individual teams, and the data we analyze 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 will archive both the analysis code used for each individual analysis and the code from our meta-analyses. We will also archive copies of our survey instruments from analysts and peer-reviewers.

We will exclude from our synthesis any individual analysis submitted after we have completed peer review or those unaccompanied by analysis files that allow us to understand what the analysts did. We will also exclude any individual analysis that does not produce an outcome that can be interpreted as an answer to our primary question.

Phase 5: Collaborative Write-Up of Manuscript

Analysts and initiating authors will discuss the limitations, results, and implications of the study and collaborate on writing the final manuscript for review as a stage-2 Registered Report.

Results

Sample description

In the following, we will describe the characteristics of our analyst teams and their models in order to give an overview over the sample and the large variation across analysis strategies.

Team properties Eighty-Four teams initially signed up to participate. Fifty-Four of the proposed teams dropped out of the project during the

analysis phase[‖]. The remaining 30 teams submitted their analysis in time and were coded by the three initiating authors (SC, JC, TR).

Submitting teams consisted of an average of 3.63 analysts (sd = 2.8). Analysts had on average 4.8 years (sd = 4.2) post PhD experience, ranging from -3.8 years, i.e. PhD students (or younger) to 12.4 years.

Analysts prior belief in the effect under investigation ranged from 48.5 to 92 with an average of 69.2 (sd = 11.3), indicating that analysts had an overall rather high prior plausibility of the investigated relationship between prosody and typicality (scale ranged from 0 to 100).

The teams have submitted 170 models to answer the research question. On average, teams submitted 3.6 models. Crucially, analytic approaches widely varied in how they acoustically analysed the speech signal and how the statistically analysed the extracted values.

Acoustic analysis Teams differed in what and how they measured the acoustic signal, including choosing different aspects of the acoustic signal, the temporal window over which they measured, and the concrete operationalization of how they measured those acoustic properties. 33% of models used a duration measure as the outcome variable, 36% used a f0 measure, 15% used a formant measure, 14% used an intensity measure, and 3% used a different measure.

45% of models measured acoustic properties at the level of the segment (e.g. comparing the acoustic profile of a vowel), 44% at the level of the word (e.g. comparing the acoustic profile of "banana"), 4% at the level of the phrase (e.g. the noun phrase including determiner and adjective, e.g. "the green banana"), 2% at the level of the whole sentence, and 5% used a different time window. Based on a coarse coding of how acoustic measures were operationalized, we find a total of 55 different measurement specifications.

Statistical analysis The multiverse of acoustic measurement choices is exponentiated with choices during the statistical analysis, including the predictors and their operationalization, the chosen inferential framework, the type of model, and their model architecture, specifically how they accounted for dependencies in the data via random effect specifications in their multilevel models.

[‖] ADD REASONS

On average, models included 1.9 different predictors (counting only conceptually different predictors), i.e. in addition to the critical predictor, typicality of the adjective noun combinations, most teams included additional predictors in their models. Predictors included the information structure of the sentence, trial number, semantic dimensions of the referent, part of speech, or gender of the speaker.

The original dataset allowed to operationalize the most relevant predictor, typicality, in different ways. It was operationalized as categorical (e.g. typical vs. atypical) in 78% of models, continuous (on a scale from 0-100) with the mean typicality rating in 19% of models, and continuous with the median typicality rating in 1% of models.

The majority of models were rooted in the frequentist framework (80%). 20% were operating in a Bayesian framework. While teams almost exclusively used linear models to analyze their data (98%, with a few exception using machine learning techniques or GAMs as a special case of a linear model), 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. The appropriate way to account for this non-independence is by using multi-level models and specifying so-called random effects (e.g., Gelman and Hill 2006; Schielzeth and Forstmeier 2009). 7% of linear models specified no random effects at all (without pooling their data), so effectively ignoring these non-independences (Hurlbert 1984). 64% specified random intercepts only, and 29% specified random intercepts and random slopes to account for the non-independence. On average, teams that specified random effects, included 2.4 random terms in their models. Based on statistical framework, type of model, distribution family, fixed terms, and discarding random effects, there were a total of 45 different model specifications.

Review ratings Teams reviewed each others' analyses for both the acoustic analysis and the statistical analysis. The mean rating of the quality of the acoustic analyses was 74.6 (sd = 11.1); the mean rating of the quality of the statistical analysis was 73.9 (sd = 14.1). For reference, the scale of the rating anchored a point grade of 75 as "an imperfect analysis but the needed changes are unlikely to dramatically alter final interpretation",

indicating a rather high confidence of reviewers that the provided analyses yield appropriate (yet "imperfect") answers to the research question.

Meta analysis

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see wh
```

Between-team variability As explained in the **Methods** section, the primary aim of this is to assess the variability of the reported effects or, in other words, 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}), i.e. the standard deviation of the group-level effect of team (`!!! we have now model_id`) returned by the meta-analytic model.

According to the meta-analytic model, the group-level standard deviation is between 0.11 and 0.15 standard units, at 95% credibility. This means that the deviations of any individual model from the meta-analytic effect estimate range between ±0.22 to ±0.3 (0.11 * 2, 0.15 * 2) in standard units. We illustrate what this means in actual acoustic measures by means of examples, taking the mean sample standard deviation of each measure from the sample data.

Estimating the crowd sourced effect size The Bayesian random effects model estimates the range of probable values of the standardized effect size between -0.01 and 0.04 (95% CrI, mean = 0.016) In other words, if we were to assume that there is a true underlying effect of typicality and we consider each analysis randomly drawn from a population of possible analysis attempts, our best guess about the underlying effect is that it is 0.02 standard deviations above zero. This outcome thus estimates that atypical word combinations have 0.02 standard deviations higher acoustic values (e.g. duration, f0 etc,) than typical word combinations.

This is not only an extremely small effect , but there is also much uncertainty around this estimate and the range of probably values includes zero. Thus, given the data, the model, and our prior assumptions, we cannot be very certain that there actually is an effect that is not zero to begin with. Since we do not know the true value of the underlying effect,

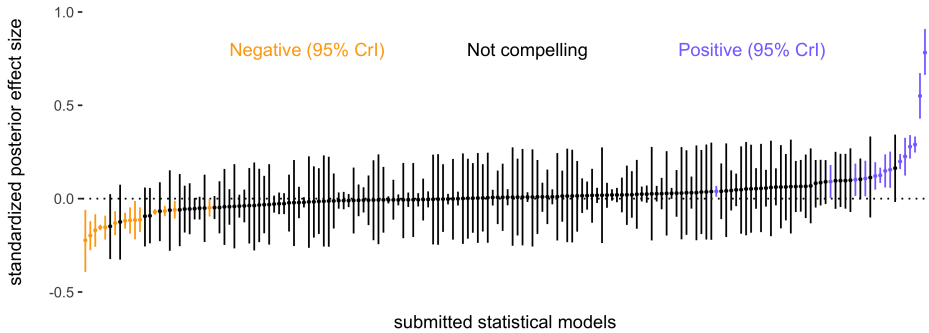

Figure 3. (ref:metaPlot1)

we cannot conclude anything from this finding, but if there is an effect of typicality, it is very small. Moreover, this population estimate is half the size of the estimated standard deviation between different models

(ref:meta_plot1) illustrates the estimated model outputs for all submitted models according to size. Given the nature and wide variety of acoustic operationalizations, there is not always a natural interpretation of the scale, but in most cases a positive effect corresponds to atypical word combinations eliciting higher acoustic values (e.g. longer duration, higher f0, etc.). Notably, while the majority of models yielded inconclusive results, there are 28 model estimates for which the 95% credible interval does not contain zero (16%).

Can we predict the estimate? After assessing the variability across models, we now turn toward estimating the impact of a series of predictors on the analysts' model estimates. There is a lot of variation, raising the question as to whether we can explain some of this variation or whether it is idiosyncratic (Breznau et al. 2021)?

(ref:meta_plot1), panel C, displays the coefficients alongside 80% and 95% credible intervals for all model predictors.

The model suggests that most team-specific predictors yield very small deviations from the meta-analytical estimate and their 95% credible intervals include zero, leaving us highly uncertain about their direction. Neither analysts' prior beliefs in the phenomenon (ADD EST + 95CrI), nor their seniority in terms of years after completing their PhD (ADD EST + 95CrI) seem to compellingly 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 (ADD EST + 95CrI). Interestingly, the model uniqueness, i.e. how unique the choice and combination of predictors is, affects the analysts estimate, with more unique models producing higher positive estimates. This suggests that the analysts' results very much depend on their choice of predictors. Looking at the most important choices during measurement, both the acoustic parameter under investigation (e.g. duration or f0) and the choice of temporal window affected the results. (ref:meta_plot1) displays 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 was measured, panel B). If an acoustic dimension related to f0 was measured, estimates are lower than the meta analytic estimate. If on the other hand, vowel formants were measured, estimates are higher than the meta analytic estimate. Similarly, if acoustic parameters were measured within individual segments (e.g. the vowel of the noun), estimates are lower than the meta analytic estimate, and if acoustic parameters were measured across whole phrases (e.g. "the blue banana"), the estimates are generally higher. In other words, depending on the choice of measurement, analysts arrived at opposite conclusions about how and if typicality is expressed acoustically.

Discussion

Summary

We gave XX analyst teams the same speech dataset to answer the same research question: *Do speakers acoustically modify utterances to signal atypical word combinations?*. In order to answer this question, teams had to operationalize latent variables within a multidimensional signal, operationalize and chose appropriate predictors, and construct an appropriate statistical model to answer the research question. As shown by the results of this meta-analytic study, such complex process has led to a wide "garden of forking paths", i.e. to a wide range of combinations of possible analytical decisions. Every individual analyst team has chosen unique paths to acoustically measure, operationalize and statistical analyse the data. Interestingly, the observed variation in reported effect sizes 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 variation in reported effect

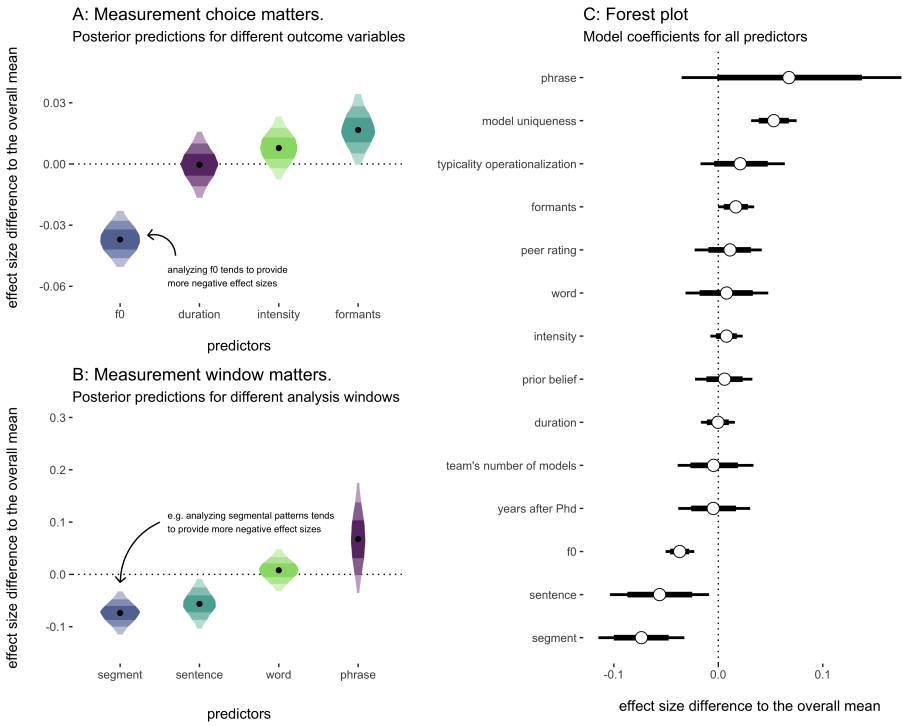

Figure 4. (ref:metaPlot2)

sizes 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)). Given the mounting evidence, Breznau et al. (2021) conclude that "[. . .] we are left to believe 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 researchers have to acknowledge and integrate into how they evaluate and present evidence.

While properties of the teams did not seem to systematically affect results, teams' estimates seem to depend on certain measurement choices. Human speech is a complex multidimensional signal. Researchers need to make choices about what to measure, how to measure it and which temporal unit to measure it in. Some outcome choices seem to bias the estimates in our data in one direction while others seem to bias estimates into another. For example, measurements related to fundamental frequency tended to result in lower estimates while measurements related to vowel formants tended to yield higher estimates. This asymmetry can have several causes. First, there could be a true underlying relationship between typicality and the speech signal that manifests itself in some measures but not others and/or manifests itself negatively in one acoustic measure but positively in another. Second, certain measurement choices might be associated with stronger expectations relative to the research question, which might lead to strong biases. Many researchers targeted measures related to voice fundamental frequency (f0) since similar functional relationships like information structure and predictability can be expressed via f0 (e.g. Grice et al. 2017; Turnbull 2017). Third, ADD REASONS. Regardless of its cause, 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 effect.

It particularly struck us that teams did not follow our instructions to only submit a single effect size. Teams submitted up to 16 different models to test for a possible relationship between typicality and the speech signal. Obviously, 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). When operating within the frequentist inferential framework, testing the same hypothesis with different dependent variables, increases the Type-I error rate if not corrected for (Tukey 1954, Benjamini & Hochberg 1995). It also invites bias, leading to the selective reporting of those tests that yield a desirable outcome (Kerr 1998, John et al. 2012, Simmon et al. 2011) while null results remain unreported (Sterling 1959, REF). Our data suggest that fields that use highly complex raw data such as speech should be particularly cautious when analyzing their data.

Lessons for the methodological reform movement

The current results point to important barriers for successful accumulation of knowledge. The replication crisis has brought attention to scientific practices that lead to unreliable and biased claims in the literature (REFS). One of the suggested paths forward is for researchers to directly replicate previous study more often [Collaboration (2015); REFS, REFS]. 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. Well-trained and experienced speech researchers in this study not only applied completely different approaches to the same research question, they also seemed to consider all these alternative approaches acceptable, as the peer-ratings suggest. Being aware of this idiosyncratic variation between analysts should lead to more nuanced claims and what Breznau et al. refers to as "epistemic humility".

A desired outcome of knowing that different but reasonable measurement choices or statistical approaches might lead to entirely 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 with 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.

In light of the observed analytic flexibility, what can we do to appropriately calibrate our confidence in our claims? First of all, through sharing of materials, data and statistical protocols, we can make our idiosyncratic choices transparent to others (REF). This enables critical evaluation and robustness checks by other fellow researchers (REF). 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 allows to extract acoustic parameters via reproducible scripts can help other researchers to trace back seemingly inconsequential choices during the measurement process [e.g., CITE Praat, EMU CITE)]].

ADD

Second, making analytic pathways completely re-traceable does not change the fact that analysts apply different analytic approaches. Crowd-source projects such as the current one can shed light on the range of degrees of freedom during analysis and help producing a consensual estimated effect using meta-analytic techniques. This 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 researchers have a good understanding of relevant analytic degrees of freedom, they could apply all conceivable analytic strategies and compare the results across all combinations of these choices. This approach is called “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 analysis is limited even further by the one group of researchers who define possible analytic pathways. Eventually, a mature scientific discipline should 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 functions are expressed in the acoustic signal. This can

be achieved by formalizing verbal expectations mathematically 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, there is potential for bias. The coordinating authors have engaged with metascientific research before and have been actively involved in methodological debates about scientific practices including transparency and statistical methods. We have in the past used the lack of standardized analytical approaches as an argument for proposing behavior and policy changes in the field. This might have biased our own judgment during the analysis which itself came with many researcher degrees of freedom. We hope we were able to make these degrees of freedom as well as the timing and reasoning of our analytical choices detectable. However, our results should be independently re-analysed and replicated using a different research question. Moreover, 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. Moreover, while the number of participating teams is larger than most earlier crowdsourcing projects, it is likely to be too small to estimate meta-analytic estimates reliably. This is particularly important for the predictor evaluation. Since predictor levels were not systematically distributed across teams, our estimates are characterized by large uncertainty. This uncertainty is possibly further inflated by the fact that he research question presented to the teams was vague: *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-analysis projects (Auspurg and Brüderl 2021). However, we consider this very underspecification of research hypotheses in the field of speech science (and beyond, see Scheel (2022)) a common phenomenon. 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). - We have not investigated researcher positionality, which is known to be an important factor.

Despite these possible sources of uncertainty and bias, our study suggest that at a minimum we can expect the between-team variability that we have observed, which is quite substantial and possibly amplifying biases.

Second, the generalizability of our findings to other disciplines on the one hand and specifically other sub-disciplines of the language sciences is, of course, limited. Our focus was on quantitative analyses that require operationalization of a multidimensional signal. While some of our results can be informative for other disciplines working with speech or video signals, they are not very informative for expected variation between analysts working in qualitative fields of research. This is an important point to make because the cognitive sciences in general, and the language sciences in particular have many research areas that are explored with qualitative methods. It is conceivable that the issues raised here to apply differently or not at all to qualitative data analysis.

Conclusion

Author contributions

See image at LINK [for peer-review, see figure `credit-taxonomy.png` attached with submission].

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.

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 .

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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